TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

---

Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

## МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

| Тема работы |
|---|
| **Алгоритмы и программное обеспечение идентификации временных конструкций в слабоструктурированных электронных медицинских текстах** |

УДК 004.421:004.415.2:004.62:61

Студент

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8ПМ9И | Журман Дмитрий Александрович | | 21.06.2021 г. |

Руководитель ВКР

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| доцент ОИТ ИШИТР | Аксенов С.В. | к.т.н. | | 21.06.2021 г. |

## КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| доцент ОСГН ШБИП | Гончарова Н.А. | к.э.н. | | |

По разделу «Социальная ответственность»

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| доцент ООД ШБИП | Антоневич О. А. | к.б.н. | | |

## ДОПУСТИТЬ К ЗАЩИТЕ:

| Руководитель ООП | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| доцент ОИТ ИШИТР | Савельев А.О. | к.т.н. | | |

Томск – 2021

**ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП**
по направлению 09.04.04 «Программная инженерия»

| Код компетенции | Наименование компетенции |
|---|---|
| **Универсальные компетенции** | |
| УК(У)-1 | Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий |
| УК(У)-2 | Способен управлять проектом на всех этапах его жизненного цикла |
| УК(У)-3 | Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели |
| УК(У)-4 | Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия |
| УК(У)-5 | Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия |
| УК(У)-6 | Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки |
| **Общепрофессиональные компетенции** | |
| ОПК(У)-1 | Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте |
| ОПК(У)-2 | Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач |
| ОПК(У)-3 | Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями |
| ОПК(У)-4 | Способен применять на практике новые научные принципы и методы исследований |
| ОПК(У)-5 | Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем |

| | |
|---|---|
| ОПК(У)-6 | Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности |
| ОПК(У)-7 | Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях |
| ОПК(У)-8 | Способен осуществлять эффективное управление разработкой программных средств и проектов |
| **Профессиональные компетенции** | |
| ПК(У)-1 | Способен к созданию вариантов архитектуры программного средства |
| ПК(У)-2 | Способен разрабатывать и администрировать системы управления базам данных |
| ПК(У)-3 | Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов |
| ПК(У)-4 | Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий |
| ПК(У)-5 | Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ |

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Савельев А.О.
(подпись) (дата) (Ф.И.О.)

**ЗАДАНИЕ**
**на выполнение выпускной квалификационной работы**

В форме:

| Магистерской диссертации |
|---|

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

| **Группа** | **ФИО** |
|---|---|
| 8ПМ9И | Журман Дмитрию Александровичу |

Тема работы:

| Алгоритмы и программное обеспечение идентификации временных конструкций в слабоструктурированных электронных медицинских текстах | |
|---|---|
| Утверждена приказом директора (дата, номер) | № 40-5/с от 09.02.2021 |

| Срок сдачи студентом выполненной работы: | 15.06.2021 |
|---|---|

**ТЕХНИЧЕСКОЕ ЗАДАНИЕ:**

| **Исходные данные к работе**<br><br>*(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).* | Объектом исследования и разработки являются алгоритмы для извлечения временных конструкций из электронных медицинских документов. |
|---|---|

| | |
|---|---|
| **Перечень подлежащих исследованию, проектированию и разработке вопросов**<br>*(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).* | 1. Обзор синтаксических парсеров.<br>2. Проектирование модуля для извлечения временных конструкций.<br>3. Проектирование модуля для поиска событий, связанных с временными конструкциями.<br>4. Тестирование.<br>5. Работа над разделом по финансовому менеджменту, ресурсоэффективности и ресурсосбережения.<br>6. Работа над разделом по социальной ответственности.<br>7. Работа над разделом на английском языке. |
| **Перечень графического материала**<br>*(с точным указанием обязательных чертежей)* | 1. Общая схема разрабатываемого ПО<br>2. Блок-схемы алгоритмов<br>3. Оценка конкурентоспособности НТИ.<br>4. Матрица SWOT.<br>5. График разработки.<br>6. Бюджет НТИ.<br>7. Диаграмма Исикавы. |

**Консультанты по разделам выпускной квалификационной работы**
*(с указанием разделов)*

| Раздел | Консультант |
|---|---|
| Основная часть | Доцент ОИТ ИШИТР, к.т.н., доцент Аксенов С.В. |
| Финансовый менеджмент, ресурсоэффективность и ресурсосбережение | Доцент ОСГН ШБИП, к.э.н., доцент Гончарова Н.А. |
| Социальная ответственность | Доцент ООД ШБИП, к.б.н., доцент Антоневич О. А. |
| Английский язык | Доцент ОИЯ, к.п.н., доцент Сидоренко Т.В. |

**Названия разделов, которые должны быть написаны на русском и иностранном языках:**

Раздел 1 Algorithms and software for time expression identification in semi-structured electronic medical texts

| | |
|---|---|
| **Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику** | 1.03.2021 |

**Задание выдал руководитель:**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| доцент ОИТ ИШИТР | Аксенов С.В. | к.т.н., доцент | | 1.03.2021 |

**Задание принял к исполнению студент:**

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8ПМ9И | Журман Дмитрий Александрович | | 1.03.2021 |

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2020 /2021 учебного года

Форма представления работы:

| Магистерская диссертация |
|---|

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

## КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
## выполнения выпускной квалификационной работы

| | |
|---|---|
| Срок сдачи студентом выполненной работы: | 15.06.2021 |

| Дата контроля | Название раздела (модуля) / вид работы (исследования) | Максимальный балл раздела (модуля) |
|---|---|---|
| 01.06.2021 | Основная часть | 70 |
| 01.06.2021 | Финансовый менеджмент, ресурсоэффективность и ресурсосбережение | 10 |
| 01.06.2021 | Социальная ответственность | 10 |
| 01.06.2021 | Английский язык | 10 |

**СОСТАВИЛ:**
**Руководитель ВКР**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| доцент ОИТ ИШИТР | Аксенов С.В. | к.т.н. | | |

**СОГЛАСОВАНО:**
**Руководитель ООП**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| доцент ОИТ ИШИТР | Савельев А. О. | к.т.н. | | |

**ЗАДАНИЕ ДЛЯ РАЗДЕЛА**
**«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»**

Студенту:

| Группа | ФИО |
|---|---|
| 8ПМ9И | Журман Дмитрию Александровичу |

| | | | |
|---|---|---|---|
| **Школа** | ИШИТР | **Отделение школы (НОЦ)** | ОИТ |
| **Уровень образования** | Магистратура | **Направление/специальность** | 09.04.04 «Программная инженерия» |

| **Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:** | |
|---|---|
| 1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих* | Работа с информацией, представленной в российских научных публикациях, аналитических материалах, и изданиях, нормативно-правовых документах. |
| 2. *Нормы и нормативы расходования ресурсов* | |
| 3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования* | |
| **Перечень вопросов, подлежащих исследованию, проектированию и разработке:** | |
| 1. *Расчет инновационного потенциала НТИ* | Оценка потенциальных потребителей исследования, SWOT-анализ. |
| 2. *Формирование плана и бюджета инженерного проекта (ИП)* | Планирование этапов работ, определение трудоемкости и построение календарного графика, формирование бюджета. |
| 3. *Оценка ресурсной, финансовой, социальной, бюджетной эффективности ИР и потенциальных рисков* | Оценка сравнительной эффективности исследования. |
| **Перечень графического материала** *(с точным указанием обязательных чертежей):* | |
| 1. Оценка конкурентоспособности НТИ<br>2. Матрица SWOT<br>3. График разработки<br>4. Бюджет НТИ<br>5. Диаграмма Исикавы | |

| | |
|---|---|
| **Дата выдачи задания для раздела по линейному графику** | 01.03.2021 |

**Задание выдал консультант:**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| доцент ОСГН ШБИП | Гончарова Наталья Александровна | к.э.н. | | |

**Задание принял к исполнению студент:**

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8ПМ9И | Журман Дмитрий Александрович | | |

# ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

| Группа | ФИО |
|---|---|
| 8ПМ9И | Журман Дмитрию Александровичу |

| | | | |
|---|---|---|---|
| **Школа** | Инженерная школа информационных технологий и робототехники | **Отделение (НОЦ)** | Отделение информационных технологий |
| **Уровень образования** | Магистратура | **Направление/специальность** | 09.04.04 «Программная инженерия» |

Тема ВКР:

| | |
|---|---|
| Алгоритмы и программное обеспечение идентификации временных конструкций в слабоструктурированных электронных медицинских текстах | |
| **Исходные данные к разделу «Социальная ответственность»:** | |
| 1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения | Объектом исследования и разработки является программная библиотека для извлечения временных конструкций из электронных медицинских документов.<br>Областью применения данной библиотеки являются предприятия, занимающиеся анализом медицинских данных.<br>Рабочей зоной является место, включающее рабочий стол и персональный компьютер. |
| Перечень вопросов, подлежащих исследованию, проектированию и разработке: | |
| **1. Правовые и организационные вопросы обеспечения безопасности:**<br>– специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства;<br>– организационные мероприятия при компоновке рабочей зоны. | - ГОСТ Р 50923-96 Дисплеи. Рабочее место оператора<br>- Федеральный закон от 27.07.2006 N 152-ФЗ (ред. от 30.12.2020) "О персональных данных"<br>- СП 2.4.3648-20 "Санитарно-эпидемиологические требования к организациям воспитания и обучения, отдыха и оздоровления детей и молодежи" |
| **2. Производственная безопасность:**<br>2.1. Анализ выявленных вредных и опасных факторов<br>2.2. Обоснование мероприятий по снижению воздействия | Вредные факторы:<br>- Отсутствие или недостаток естественного света.<br>- Недостаточная освещенность рабочей зоны.<br>- Перенапряжение анализаторов.<br>- Превышение уровня шума на рабочем месте.<br>- Повышенный уровень электромагнитных излучений.<br>- Статические перегрузки, связанные с рабочей позой.<br>Опасные факторы:<br>- Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека. |

| | |
|---|---|
| **3. Экологическая безопасность:** | - Воздействие объекта на литосферу: утилизация электротехнических компонентов и отходов (вышедший из строя ПК, люминесцентные лампы). |
| **4. Безопасность в чрезвычайных ситуациях:** | - Возможной ЧС при разработке и эксплуатации проектируемого решения является возникновение пожара. |

| | |
|---|---|
| **Дата выдачи задания для раздела по линейному графику** | 01.03.2021 |

**Задание выдал консультант:**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| доцент ООД ШБИП | Антоневич Ольга Алексеевна | к.б.н. | | |

**Задание принял к исполнению студент:**

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8ПМ9И | Журман Дмитрий Александрович | | |

## Реферат

Работа содержит пояснительную записку на 105 листах, содержит 8 рисунков, 25 таблиц, 3 приложения, 47 источников.

Ключевые слова: синтаксический парсинг, временная конструкция, Spacy, DeepPavlov, электронные медицинские карты.

Цель данной работы – повышение эффективности анализа электронных медицинских карт (ЭМК) с помощью разработки инструментов автоматического извлечения временных конструкций из медицинской документации. Полученные инструменты позволят перенести данные конструкции на временную шкалу и представить их в удобном для медицинских сотрудников виде. Также результаты работы могут быть востребованы научными центрами, занимающимися применением машинного обучения в области медицины, так как полученные инструменты позволяют собрать необходимые для обучения данные.

**Содержание**

**Обозначения и сокращения**

В данной работе используются следующие определения и обозначения:

ПО – программное обеспечение;

ВК – временная конструкция;

ЭМК **–** электронные медицинские карты;

SVM – support vector machine;

LSTM – Long short-term memory;

GRU – Gated Recurrent Units;

CRF – Conditional random field.

## Введение

Во время заболевания или нахождения в больнице с пациентом происходят различные события. Например, пациент получает лечение, появляются новые симптомы, выполнятся операции и т.д. Если переложить вышеперечисленные события из анамнеза на временную шкалу, врач сможет наглядно видеть, что и когда происходило с пациентом. Также это позволит подготовить набор данных для предиктивных моделей в медицине и здравоохранении. Информацию о данных событиях можно получить из анамнеза Электронных Медицинских Карт. К сожалению, анамнез пациента имеет неструктурированную форму. Усложняет задачу то, что для русского языка отсутствует размеченный набор медицинских данных. Следовательно, решение данной задачи возможно только при помощи синтаксического и морфологического анализа, а также грамматических правил.

Таким образом, цель работы – повышение эффективности анализа электронных медицинских карт с помощью разработки инструментов автоматического извлечения временных конструкций из медицинской документации. Полученные инструменты позволят перенести данные конструкции на временную шкалу и представить их в удобном для медицинских сотрудников виде. Также результаты работы могут быть востребованы научными центрами, занимающимися применением машинного обучения в области медицины, так как полученные инструменты позволяют собрать необходимые для обучения данные. Теоретическая значимость данной работы заключается в разработке алгоритмов и методов по извлечению значимой информации из ЭМК, не прибегая к методам машинного обучения.

Для достижения цели были поставлены следующие задачи:

1. Обзор и сравнение синтаксических парсеров для русского языка;
2. Разработка модуля для извлечения временных конструкций;
3. Разработка модуля для поиска событий, связанных с временными конструкциями.

# 1. Системы извлечения временных конструкций из медицинских текстов

На данный момент состоялось несколько соревнований по извлечению временных конструкций из медицинских текстов: I2B2 Temporal Challenge (2013), Clinical TempEval Shared Task (2015, 2016 и 2017). Во всех соревнованиях требовалось извлечь временные конструкции, события и отношения между ними, используя аннотированный корпус текстов [1].

Решения, основанные на Clinical TempEval Shared Task, использовали методы обучения с учителем для решения проблемы. Чаще всего применялись рекуррентные нейронные сети (LSTM / GRU) [2–5]. В более ранних статьях использовались модели SVM [6,7], а в более современных работах используются attention models [8,9]. В некоторых случаях применялись сверточные нейронные сети [10] или модель BERT для извлечения временных событий [11]. В любом случае мы не можем использовать методы, описанные в этих работах, для решения проблемы, так как они использовали для обучения аннотированный корпус для английского языка, а для русского языка такой корпус отсутствует.

В I2B2 Temporal Challenge участники использовали комбинированные методы, то есть совмещали обучение с учителем и методы, основанные на правилах. Грамматические правила использовались для поиска временных конструкций и отношений между событиями. Также применялся синтаксический анализ для поиска потенциальных событий. Наконец, события извлекались с помощью модели SVM [1,12].

Методы, примененные в I2B2 Temporal Challenge, представляют больший интерес для данной работы, так как участники соревнований полагались не только на обучение с учителем. Поэтому в данном разделе приведен обзор решений I2B2 Temporal Challenge.

## 1.1 Гибридная система извлечения временной информации из клинического текста

Данная система заняла первое место в I2B2 Temporal Challenge и достигла F-score 0,8659 для излечения временных конструкций. Для извлечения временных конструкций в данном решении используется основанная на грамматических правилах система. Большинство правил было заимствовано из системы HeidelTime. В предложении «She was admitted on the morning of Feb 25, 2009.» система сначала извлекает «the morning of Feb 25, 2009» как временную конструкцию со значением «Feb 25, 2009» и словом-модификатором «morning», используя правило (% PartOfDate) (% WordMonth% DigitDay,% FourDigitYear) ', где PartofDate - это словарь всех терминов, представляющих части дат (например, «morning», «afternoon» и «evening»), WordMonth - это словарь всех английских слов, обозначающих месяцы (например, «Jan», «February»), DigitDay - это словарь номеров календарного дня (например, 1–31), а FourDigitYear - это регулярное выражение для четырехзначных номеров года (например, «1984», «2009»). Затем значение «Feb 25, 2009» нормализуется до «2009-02-25». Наконец, временная конструкция извлекается следующим образом: TIMEX3 = ' the morning of Feb 25, 2009 ' || type = 'DATE' || val = '2009-02-25'.

Затем для поиска потенциальных событий, связанных с временными конструкциями, использовался Стэнфордский синтаксический парсер. Таким образом, в качестве событий рассматривалось каждое последовательное словосочетание возле временной конструкции или любое словосочетание, имеющее связь с временной конструкцией на основе синтаксического анализа зависимостей предложения. Для финального определения события использовалась обученная SVM модель [1].

### 1.2 Комплексное обнаружение временной информации из клинического текста: медицинские события, время и идентификация TLINK

Данная система показала лучший результат по извлечению временных конструкций среди участников I2B2 Temporal Challenge и достигла F-score 0,9000.

Для извлечения временных конструкций в данном решении используется основанная на регулярных выражениях система. Большинство правил было заимствовано из системы HeidelTime. Извлеченные события имеют три основных атрибута: тип (дата, время, продолжительность, частота), значение (нормализованная форма) и модификатор (приблизительно, больше, меньше, начало, конец, середина).

Затем для поиска потенциальных событий, связанных с временными конструкциями, использовалась система icTAKES. Данная система выполняет синтаксический анализ предложений, токенизацию, нормализацию терминов, тегирование части речи, разбиение на части, обнаружение отрицания и разрешение неопределенности. Также для поиска событий использовались разработанные правила. Например, результаты анализов всегда предшествуют постановке диагноза.

Таким образом, в качестве потенциальных событий рассматривалось последовательное словосочетание, удовлетворяющее разработанным правилам, или любое словосочетание, имеющее связь с временной конструкцией на основе синтаксического анализа зависимостей предложения. Для финального определения события использовались CRF и SVM модели [12].

### 1.3 Сквозная система для определения временных отношений в выписках

Данная система смогла достичь F-score 0,8818 в I2B2 Temporal Challenge.

Для извлечения событий в данном решении используется основанная на двух CRF моделях. В качестве признаков для модели использовались:

нормальная форма слова, N-символьный префикс и суффикс, кластеризация слов, первые и последние восемь букв каждого слова.

Для извлечения временных конструкций данная система также использует CRF модель. Нормализация извлеченных ВК осуществляется при помощи алгоритма контекстно-свободной грамматики и нескольких multi-SVM классификаторов. Для некоторых ВК нормализация может быть выполнена непосредственно из текста предложения или текста временной конструкции, например, «2012-10-10». Однако есть некоторые ВК, нормальная форма которых должна быть выведены из всего документа, например, «пятый послеоперационный день». Данная система нормализации состоит из двух этапов. Сначала нормализация осуществляется только при помощи информации, полученной из текста предложения и ВК. Временные конструкции, нормальная форма которых не может быть получена из текста предложения, преобразовываются в промежуточные представления. На основе промежуточного представления составлено несколько правил, которые позволяют получить конечную нормальную форму ВК.

Для получения конечного результата данные полученные при извлечении ВК и событий обрабатываются при помощи десяти multi-SVM классификаторов и одной логико-марковской сети.

Таким образом, для извлечения событий и извлечения ВК использовались CRF модели. Для нормализации применялся алгоритм контекстно-свободной грамматики и нескольких multi-SVM классификаторов. Если нормальную форму не удавалось получить из текста предложения, использовались составленные правила. Для финального определения события и ВК применялись десять multi-SVM классификаторов и одна логико-марковской сеть [13].

### 1.4 Выводы

Из приведенного выше обзора методов, примененных в I2B2 Temporal Challenge следует:

1. Для извлечения временных конструкций можно достичь F-score 0,85 и выше, используя методы, основанные на грамматических правилах.

2. В качестве потенциальных событий, связанных с временными конструкциями, можно использовать любое словосочетание, имеющее связь с ВК на основе синтаксического анализа зависимостей предложения.

Данные подходы будут использоваться при дальнейшей разработке алгоритмов и программного обеспечения идентификации временных конструкций в слабоструктурированных электронных медицинских текстах.

# 2. Проектирование программного обеспечения идентификации временных конструкций

В данной главе приводится описание основных компонентов и проектируемого программного обеспечения идентификации временных конструкций, используя подходы, полученные в предыдущей главе.

## 2.1 Архитектура разрабатываемого программного обеспечения

На рисунке 1 показаны основные части разрабатываемого программного обеспечения.

Для поиска событий, связанных с временными конструкциями, требуется выполнить синтаксический анализ зависимостей предложения. Обучение собственного синтаксического парсера требует размеченный корпус и большие вычислительные мощности. Поэтому в данной работе будет использоваться уже обученный синтаксический парсер.

Также для достижения цели необходимо реализовать извлечение временных конструкций из предложений с помощью методов, основанных на грамматических правилах. Эти конструкции следует нормализовать, то есть привести к единому формату (ГГГГ-ММ-ДД). На данный момент уже существует ряд готовых библиотек (dateparser и rutimeparser [14,15]) для извлечения временных конструкций и их нормализации. Однако при работе с медицинским текстом они показывают низкую точность. Это связано с тем, что корпус медицинских текстов содержит повторяющиеся события и имеет множество форматов временных конструкций. Поэтому необходимо разработать собственный модуль, основанный на грамматических правилах, для извлечения временных конструкций из предложений.

После этого в синтаксическом дереве нам нужно построить путь от найденной временной конструкции до события, к которому оно принадлежит.

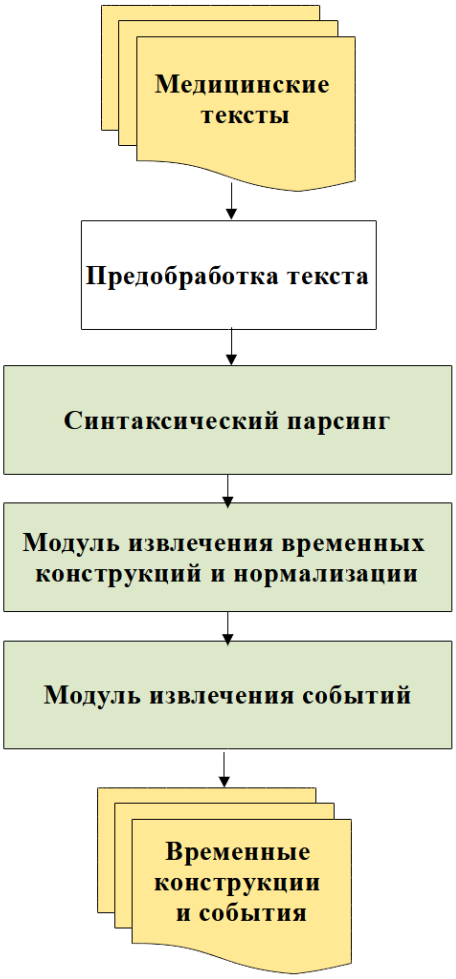


Рисунок 1 – Общая схема разрабатываемого ПО

## 2.2 Выбор синтаксического парсера для русского языка

Каждый год проводятся соревнования среди парсеров CoNLL Shared Task. Однако в 2019 году соревнования проводились только для английского языка, а в 2020 году соревнования не состоялись [16]. Результаты за 2018 год представлены в таблице 1 [17]. На CoNLL Shared Task сравнивается, как парсеры работают с русским языком на корпусе syntagrus. Данный корпус содержит 61889 предложений и 1106296 токенов. Все предложения имеют синтаксическую разметку, выполненную лингвистами в формате Universal Dependencies [18].

Помимо CoNLL Shared Task проводились соревнования по полной грамматической разметке текстов на русском языке GramEval-2020. Результаты данных соревнований для синтаксического парсинга представлены в таблице 2 [19].

Стоит отметить, что корпус и парсеры постоянно обновляются, поэтому было проведено повторное сравнение последних версий парсеров на последней версии корпуса syntagrus. Помимо точности парсеров необходимо определить, какие парсеры чувствительны к пунктуации, так как врачи в анамнезе допускают много пунктуационных ошибок.

Также можно выделить следующие популярные синтаксические парсеры для русского языка, которые не принимали участия в CoNLL Shared Task и в GramEval-2020, но будут рассмотрены в данной работе: Slovnet, Deep Pavlov и Spacy ввиду своей популярности [20–22].

Таблица 1 – Результаты CoNLL 2018 Shared Task

| Парсер | LAS |
|---|---|
| HIT-SCIR (Harbin) | 92.48 |
| TurkuNLP (Turku) | 91.72 |
| Stanford (Stanford) | 91.59 |
| UDPipe Future (Praha) | 91.46 |
| ParisNLP (Paris) | 91.41 |

Таблица 2 – Результаты GramEval-2020 (метрика LAS)

| | Fiction | News | Poetry | Social | Wiki | 17 cent |
|---|---|---|---|---|---|---|
| qbic | 0.896 | 0.912 | 0.814 | 0.807 | 0.781 | 0.665 |
| ADVance | 0.869 | 0.911 | 0.780 | 0.784 | 0.760 | 0.618 |
| lima | 0.850 | 0.843 | 0.725 | 0.713 | 0.697 | 0.546 |
| vocative | 0.826 | 0.834 | 0.660 | 0.659 | 0.694 | 0.500 |
| Turku | 0.859 | 0.877 | 0.731 | 0.733 | 0.711 | 0.502 |
| Stanford | 0.854 | 0.873 | 0.709 | 0.706 | 0.703 | 0.509 |
| UDPipe | 0.811 | 0.817 | 0.666 | 0.644 | 0.668 | 0.462 |
| SyntaxNet | 0.808 | 0.802 | 0.6 | 0.614 | 0.645 | 0.446 |
| MaltParser | 0.599 | 0.553 | 0.404 | 0.476 | 0.436 | 0.340 |

Для сравнения синтаксических парсеров используется метрика UAS (Unlabeled attachment score), которая оценивает корректно ли определена вершина слова, и LAS (Labeled attachment score), которая оценивает правильно ли найдена вершина, а также тип зависимости. Для вычисления данных метрик использовался модифицированный скрипт с соревнований CoNLL Shared Task [23]. Сравнение производилось на приватной части корпуса syntagrus (6491 предложений).

Стоит отметить, что парсеры запускались на производительной рабочей станции (RAM - 64GB, GPU - 2080Ti, CPU - ryzen 3900 с 12 ядрами), поэтому скорость работы может отличаться при запуске на обычных ПК.

Результаты сравнения приведены в Таблице 3. Как можно заметить из таблицы, лучшую точность по метрике LAS показало решение qbic, несмотря на то, что оно не обучалось на корпусе syntagrus. Также стоит отметить неплохую точность парсера DeepPavlov. Однако парсер qbic является решением для конкретных соревнований GramEval-2020, а не готовым пакетом для какой-нибудь задачи в отличие от Stanza, UDPipe, DeepPavlov и Slovnet. Если внимательно изучить репозитории таких парсеров с соревнований CoNLL 2018 Shared Task, то можно заметить, что данные решения не обновлялись с момента окончания соревнований. Поэтому в данной работе в дальнейшем будет использоваться синтаксический парсер DeepPavlov, так как он регулярно дорабатывается и содержит подробную документацию.

Таблица 3 – Результат сравнения парсеров

| Парсер | Скорость работы (sent/s) | UAS | LAS | Чувствителен к пунктуации |
|---|---|---|---|---|
| qbic | 55 (GPU) | 96.98 | 94.16 | - |
| DeepPavlov | 53 (GPU) | 93.40 | 92.07 | - |
| Stanza | 18 (GPU) | 91.75 | 89.97 | + |
| Turku | 21 (CPU) | 88.36 | 87.08 | + |
| UDPipe | 87 (GPU) | 87.67 | 80.06 | + |
| Slovnet | 705 (CPU) | 81.05 | 76.43 | - |
| Spacy | 40 (CPU) | 86.93 | 75.56 | + |

### 2.3 Выбор парсера контекстно-свободных грамматик для русского языка

Среди парсеров контекстно-свободных грамматик для русского языка можно выделить Yargy, Spacy и Tomita [22,24,25]. Парсер необходимо выбрать исходя из следующих требований:

- Скорость работы.
- Наличие подробной документации.

- Возможность написания правил на основном языке разрабатываемого ПО – python.

Tomita – разработка компании Яндекс. Библиотека реализована на языке C++. Однако Tomita парсер для описания правил использует свой язык, а в Yargy и Spacy все правила описываются на языке Python. Также в Tomita парсере все правила и грамматики являются закрытыми, а Yargy и Spacy представляют собой проект с исходным кодом, доступным на Github. Стоит отметить, что скорость работы Tomita парсера превосходит скорость Yargy и Spacy, благодаря тому, что он реализован на C++ [24,25].

Yargy использует морфологический анализатор Pymorphy, который также является библиотекой с открытым исходным кодом. Более того, у него есть несколько разработанных готовых наборов правил для извлечения таких сущностей, как адреса, имена, даты и др., которые доступны в репозитории проекта Natasha на Github[24].

Spacy – это бесплатная библиотека с открытым исходным кодом для обработки естественного языка в Python. В ней имеется возможность поиска именованных сущностей, определения частей речи. Однако в Spacy отсутствует морфологический анализатор для русского языка. В данной работе это не является проблемой, так как синтаксический парсер DeepPavlov способен выполнять морфологический анализ [25].

Для сравнения скорости работы парсеров был написан набор из 100 правил для Spacy и для Yargy. Spacy удалось достичь скорости около 35 предложений в секунду. Скорость обработки yargy составила всего 7 предложений в секунду. Такая разница в скорости работы возникла из-за того, что у Spacy имеется возможность работать на GPU. Помимо этого, yargy повторно осуществляет морфологический анализ с помощью Pymorphy после того, как все предложения уже были обработаны парсером DeepPavlov. Spacy не осуществляет повторную обработку, а сразу использует результаты работы DeepPavlov.

Таким образом, в дальнейшей работе будет использовать парсер Spacy из-за его высокой скорости и возможности написания правил на языке python.

## 2.4 Выводы

Таким образом, для дальнейшей работы будет использован синтаксический парсер DeepPavlov. Среди других парсеров его выделяет скорость работы, подробная документация и точность определения связей.

Также в качестве основного парсера контекстно-свободных грамматик для русского языка будет использован Spacy. Это позволит писать правила на языке Python. Также скорость работы Spacy в пять раз выше, чем у парсера Yargy.

# 3. Разработка алгоритмов и программного обеспечения идентификации временных конструкций

В данной главе приводятся основные этапы разработки алгоритмов и программного обеспечения идентификации временных конструкций и описание проблем, возникнувших в ходе работы.

## 3.1 Создание правил контекстно-свободных грамматик для поиска временных конструкций и нормализации

Правила в парсере Spacy выглядят следующим образом:

{"LOWER": "привет"}, {"IS_PUNCT": True}, {"LOWER": "мир"}

В фигурных скобках заключается название атрибута и его значение. Список основных атрибутов приведен в таблице 4 [22].

Таблица 4 – Список основных атрибутов Spacy

| Атрибут | Описание |
|---|---|
| TEXT | Точный дословный текст токена. |
| LOWER | Текст токена в нижнем регистре. Все заглавные буквы в результате преобразования станут строчными |
| LENGTH | Длина текста токена. |
| IS_ALPHA, IS_ASCII, IS_DIGIT | Текст токена состоит из буквенных символов, символов ASCII или цифр. |
| IS_LOWER, IS_UPPER, IS_TITLE | Текст токена написан строчными, прописными или заглавными буквами. |
| IS_PUNCT, IS_SPACE, IS_STOP | Токен является знаком препинания, пробелом ли стоп-словом. |
| IS_SENT_START | Токен - это начало предложения. |
| LIKE_NUM, LIKE_URL, LIKE_EMAIL | Текст токена напоминает номер, URL-адрес, адрес электронной почты. |
| POS, TAG, DEP, LEMMA | Простой и расширенный тег части речи токена, метка зависимости, лемма. |
| ENT_TYPE | Метка сущности токена |
| _ | Атрибуты, определенные пользователем |

Пример правила для временной конструкции вида «10 лет назад» приведен ниже:

{"IS_ DIGIT ": True}, {"LEMMA": {"IN": ['день', 'час', 'неделя', 'год', 'месяц']}}, {"TEXT": "назад"}

Далее извлеченные временные конструкции необходимо было нормализовать. Для каждого из правил была написана лямбда функция для преобразования даты к формату ГГГГ-ММ-ДД при помощи стандартной библиотеки datetime языка Python и библиотеки dateutil.

Пример лямбда функции для нормализации временной конструкции вида «месяц назад» приведен ниже:

```
lambda ent: ent.doc._.date-relativedelta(**{relative_dict[ent[0].lemma_]:1})
```

Для правильной нормализации необходимо знать дату осмотра. Она хранится в переменной ent.doc._.date. Затем из даты осмотра нужно отнять интервал времени равный одному месяцу. Длины интервалов времени для каждого отрезка времени (месяц, год, неделя, день, час) хранятся в словаре relative_dict. Лямбда функция определяет по первому слову из временной конструкции длину интервала времени из словаря и вычитает из даты осмотра.

Также каждой временной конструкции присваивается метка:

- 1 – разовая ВК. Например, месяц назад или вчера;
- 2 – продолжительная ВК. Например, на протяжении 2 лет, последние 2 месяца;
- 3 – повторяющаяся ВК. Например, каждый месяц, 2 раза в год;
- 4 – зависимая ВК. Например, через месяц после выписки, спустя год после ремиссии.

Таким образом, для работы модуля было написано свыше 250 правил. Общее правило в разрабатываемом ПО выглядит следующим образом:

```
'r_halfyear_ago': {'pattern': [{"LEMMA": "полугода"}, {" LEMMA ": "назад"}]
                'norm': lambda ent: ent.doc._.date-relativedelta(months=6),
                'stamp': 1}
```

Где:

- 'r_halfyear_ago' – уникальное имя правила;
- 'pattern' – Spacy правило для извлечения ВК;
- 'norm' – лямбда функция для нормализации ВК;
- 'stamp' – метка.

### 3.2 Алгоритм поиска событий, связанных с временными конструкциями

В основе модуля лежит алгоритм с рекурсией, приведенный на рисунке 2. То есть, если на вход модуля подается сложное предложение, модуль разбивает его на части. Если ВК находится в первой части предложения, то алгоритм работает с данной частью как с обычным простым предложением, игнорируя другие части. Если ВК находится не в первой части, модуль убирает из предложения первую часть и заново определяет, где находится ВК. Так продолжается до тех пор, пока ВК не окажется в первой части предложения или не останется всего одна часть. Например, в предложении «Постоянно беспокоит головокружение, летом 2010 года однократно эпизод потери сознания» модуль после того, как не смог найти ВК в первой части, отбрасывает ее, а затем работает с частью «летом 2010 года однократно эпизод потери сознания» как с обычным простым предложением. На рисунке 3 показан результат синтаксического анализа этого предложения. Во второй части предложения нет слов в древе, образованном временным выражением. Также в древе восемь слов и четыре ветви. Помимо этого, нет ветви с типом зависимости «номинальный субъект». Это означает, что алгоритм извлек данные из крайней правой ветви и получит результат «эпизод потери сознания».

После этого была выполнена оценка точности на первых 500 предложениях. При оценке по метрике accuracy точность составила 53,8%. Однако данная метрика чувствительна к порядку слов и пунктуации, поэтому был проведен подробный анализ ошибок, приведенный в таблице 5. Из анализа можно заметить, что самой частой ошибкой для метрики accuracy была «спорная разметка». Это связано с тем, что разметка была выполнена неидеально, то есть иногда два события в предложении относятся к одной временной конструкции. Таким образом модуль извлекал одно событие, а в разметке было указано другое. Также иногда модуль включал в текст события дополнение, которое не было указано в разметке, или наоборот не находил

дополнение, которое было в разметке. Таким образом, все 97 предложений, относящиеся к данной категории на самом деле не являются ошибками.

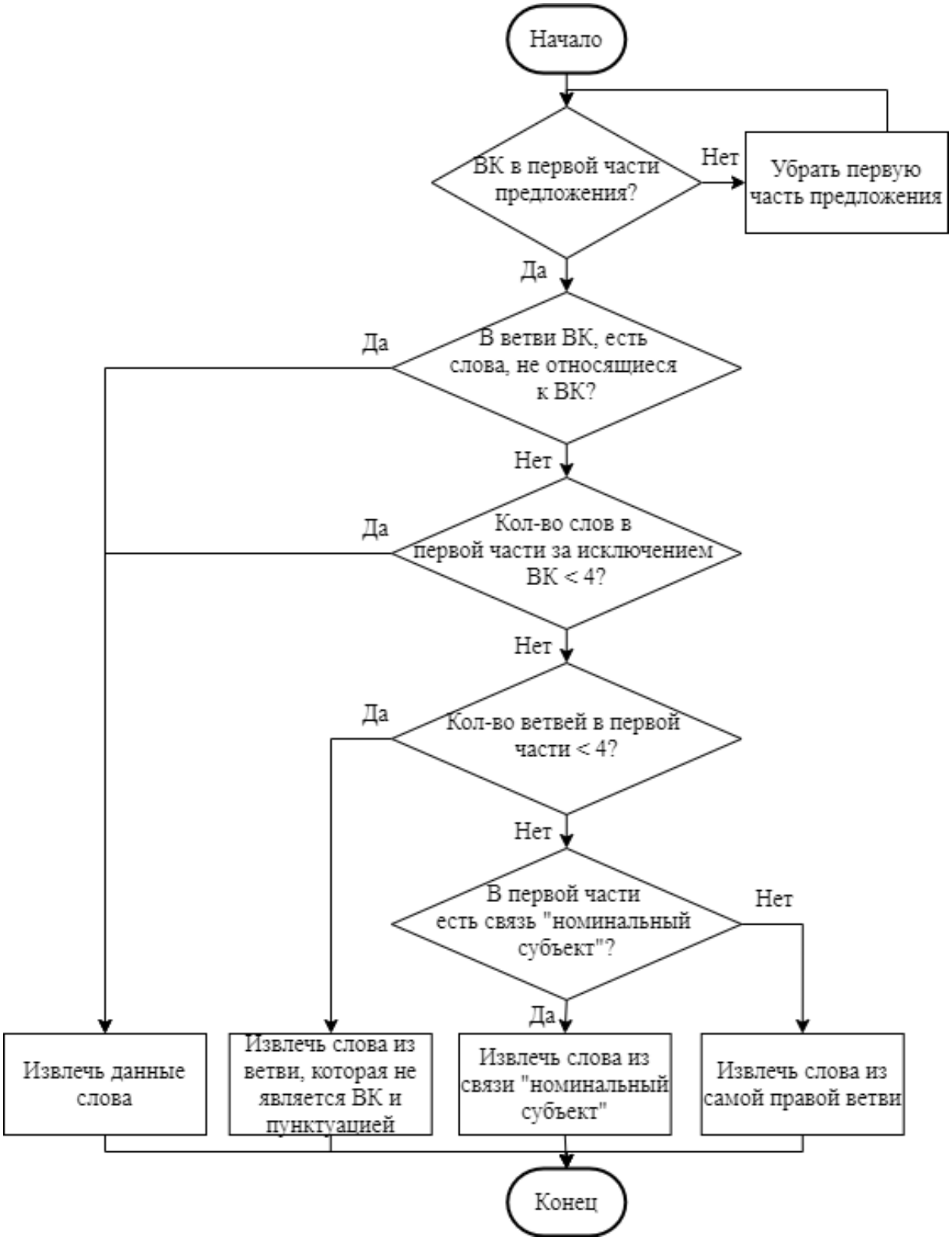


Рисунок 2 – Алгоритм работы модуля для извлечения событий

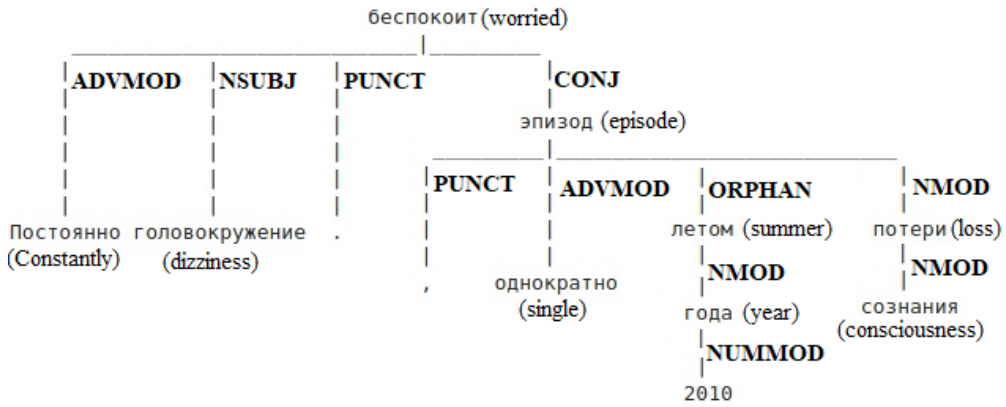


Рисунок 3 – Синтаксическое древо предложения

Таким образом, если не считать различный порядок слов и спорную разметку, точность извлечения событий составит 75%.

Таблица 5 – Анализ ошибок при извлечении событий

| **Причина ошибки** | **Количество** |
|---|---|
| Спорная разметка | 97 |
| Ошибка синтаксического парсера | 28 |
| Ошибка при извлечении события | 26 |
| Ошибка при извлечении временной конструкции | 28 |
| Найдены не все однородные члены | 13 |
| Различный порядок слов | 9 |
| Пропущен предлог | 9 |
| Орфографическая ошибка в предложении | 8 |
| Пропущена частица «НЕ» | 5 |
| В предложении 2 события, относящиеся к 2 разным временным конструкциям в 1 ветке | 4 |
| Ошибка при извлечении событий и некорректно составленное предложение | 3 |
| Пропущено тире | 2 |
| Спорная разметка, некорректно составленное предложение | 1 |
| Грамматическая и орфографическая ошибка в предложении | 1 |
| Извлечены не все слова, относящиеся к событию | 1 |
| Ошибка синтаксического парсера и некорректно составленное предложение | 1 |
| Часть события находится в придаточном предложении | 1 |
| Грамматическая ошибка в предложении | 1 |
| Ошибка синтаксического парсера и ошибка при извлечении ВК | 1 |
| Сложная грамматическая конструкция | 1 |
| Ошибка синтаксического парсера из-за наличия слов на латинице | 1 |

Согласно результатам эксперимента, самой частой причиной ошибок является алгоритм модуля извлечения событий. Так как в данной работе мы имеем дело с естественным неструктурированным языком, невозможно

разработать алгоритм, который бы подходил под все предложения естественного языка. При анализе ошибок были обнаружены предложения, которые противоречат алгоритму и их исправление ведет к еще большему возникновению ошибок.

В некоторых случаях происходили ошибки в синтаксическом парсере. Как было указано выше, из-за отсутствия размеченного набора медицинских текстов, с данной категорией ошибок невозможно что-то сделать.

Помимо этого, были обнаружены случаи, когда модуль находит не все однородные члены, принимая их за придаточное предложение.

### 3.3 Разработка модуля поиска отрицаний

В некоторых случаях парсер совершает серьезную ошибку и извлекает событие правильно, но без частицы НЕ или частицы БЕЗ, что кардинально меняет смысл. Это связано с тем, что в синтаксическом древе частицы НЕ и БЕЗ иногда находятся в другой ветке от события, и в русском языке для них нет особого типа связи в отличии от английского языка. Необходимо не только правильно определить отрицательную частицу, но и найти отрицаемую сущность. Это позволит понять, входит отрицание в ВК или нет. Решение данной проблемы поможет увеличить точность работы модуля. Также это позволит разработать дополнительный инструмент обработки медицинских документов.

Для исправления данной ошибки использовался синтаксический парсинг. Для поиска наиболее часто встречаемых шаблонов отрицаний использовался парсер Spacy.

Примеры наиболее часто встречаемых шаблонов отрицаний:

- Дополнений нет;
- Отрицает «название заболевания»;
- «Название заболевания» ранее отрицает;
- Без лекарственного покрытия.

В основе модуля поиска отрицаний лежит алгоритм, приведенный на рисунке 4. На вход подается предложение с отрицанием. Если отрицательная частица – это слова «отрицать», «отказаться» или «нет», из древа, образованном отрицательной частицей извлекаются слова из ветки со связью «номинальный субъект».

После этого идет проверка, есть ли во всем древе связь с типом «номинальный субъект». В таком случае извлекаются все слова из данной ветки.

Если предложение сложное, то модуль определяет часть, в которой находится отрицательная частица, и продолжает работать с ней как с обычным предложением.

Если вершиной древа является существительное, то извлекается вершина, дочернее слово вершины и отрицательная частица.

Если в древе меньше четырех ветвей, то извлекаются все слова кроме пунктуации. Если ни одна из проверок не прошла, то извлекается только вершина древа и отрицательная частица.

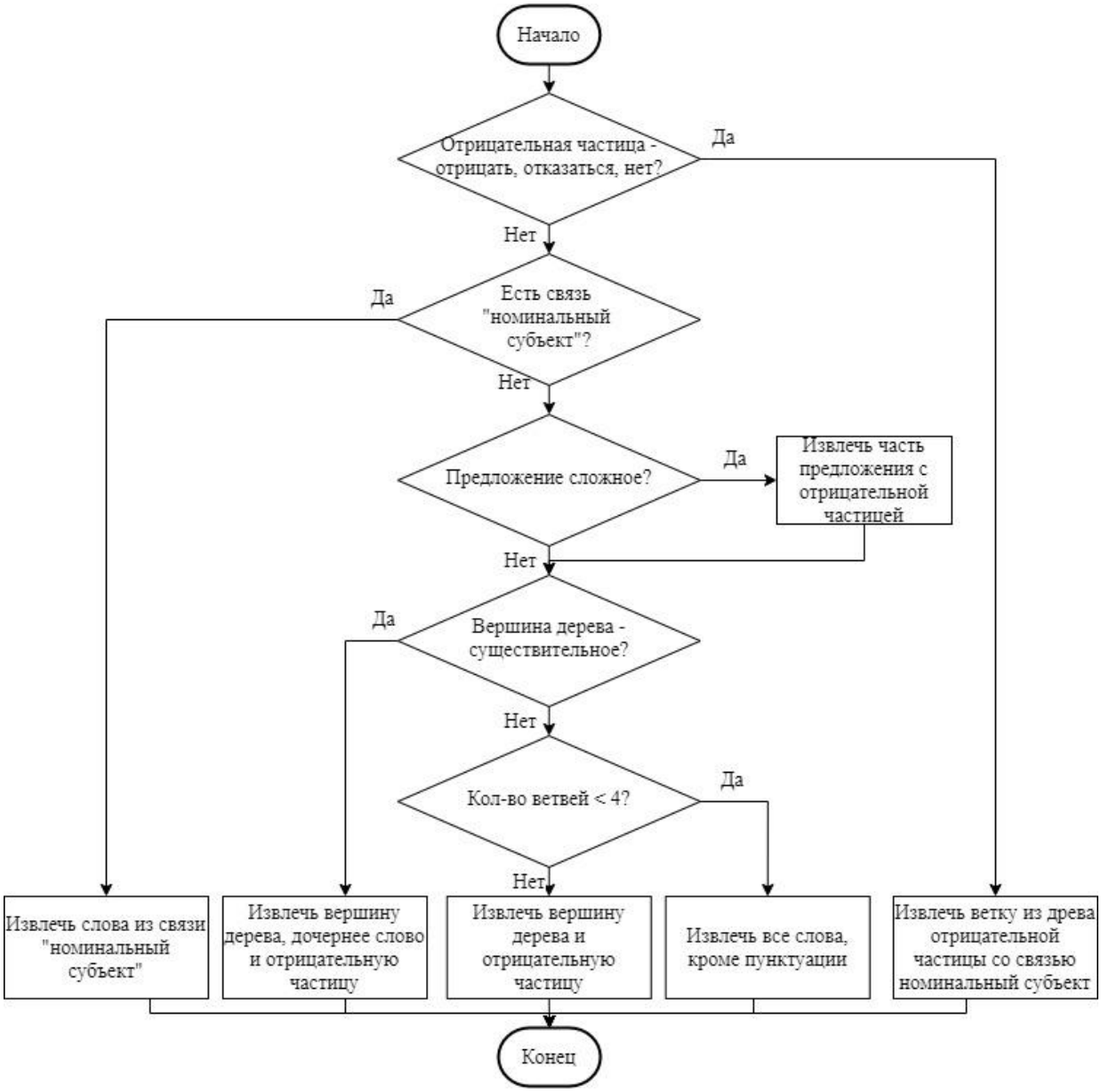


Рисунок 4 – Алгоритм работы модуля для отрицаний

После проверки модуля на размеченном наборе данных удалось достичь точности работы 78,4%. Встраивание данного алгоритма поиска отрицаний в разрабатываемое программное обеспечение позволит точнее определять события, связанные с ВК.

### 3.4 Обработка неопределенностей

Одна из проблем, возникающих при работе с временными конструкциями – это неопределенность. Например, когда в карте пациента появляется выражение «10 лет назад», это не означает, что событие произошло

ровно 10 лет назад. Необходимое событие может находиться в определенном диапазоне вокруг этой даты.

Можно выделить следующие типы нечеткости [26]:

- Неопределенность. Некоторая временная информация неизвестна или неточна. Например, «Диабет 2 типа был диагностирован около 5 лет назад»;
- Субъективность. Временные события или периоды могут быть определены субъективно или неоднозначно. Например, «После операции инфарктов не было»;
- Неопределенность. События могут быть определены с различной степенью детализации. Например: «Днем нет эпизодов сонливости».

Можно было бы смоделировать неопределенность «10 лет назад» как функцию плотности вероятности во времени с наибольшей вероятностью около 10 лет назад (например, гауссова функция), но вычисление выводов на основе такого представления NP-сложно [27].

Подход, реализованный в [28], моделирует неопределенность в виде интервала с расширяющимися границами. Этот метод представляет выражение, содержащее неопределенность, как плюс-минус одна временная единица (день, неделя, месяц или год) от точки на временной линии, к которой это событие относится. Например, выражение «Сыпь появилась за три недели до поступления» было представлено как интервал от двух до четырех недель. Модификаторы неопределенности, такие как «около» и «примерно», могут расширять интервал: с -50% до + 100%. То есть выражение «Около трех недель» можно представить, как интервал от 1,5 до шести недель. Размер диапазона может зависеть от контекста.

Другой подход к представлению неопределенности – это нечеткая логика. В теории нечетких множеств, мы можем использовать функцию принадлежности *I*, которая представляет наш уровень уверенности в том, что событие *t* находится в четком интервале *i*. Если $I(t) = 0$, мы полностью уверены, что *t* не входит в *i*; если $I(t) = 1$, мы полностью уверены, что *t*

находится в *i* [26]. Функцию принадлежности можно создать несколькими способами [29].

Треугольная функция принадлежности:

$$f(x,a,b,c) = \max(mix(\frac{x-a}{b-a},1,\frac{c-x}{c-b}),0) \quad (1)$$

Трапециевидная функция принадлежности:

$$f(x,a,b,c,d) = \max(\min(\frac{x-a}{b-a},1,\frac{d-x}{d-c}),0,) \quad (2)$$

Рисунок 5 демонстрирует треугольную и трапециевидную функции принадлежности.

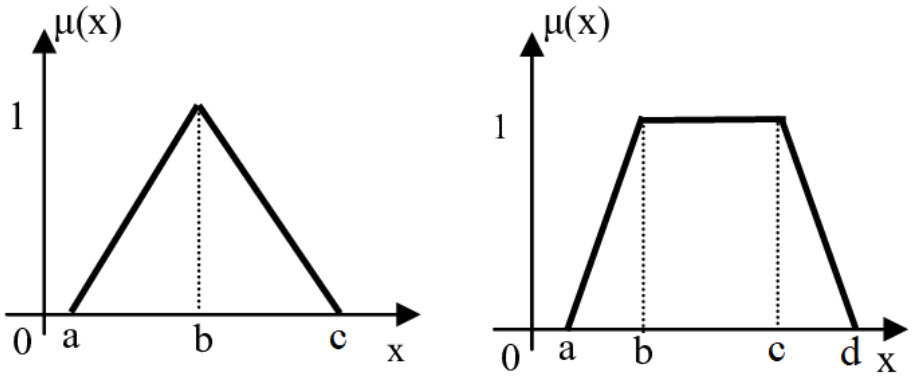


Рисунок 5 – Треугольная и трапециевидная функции принадлежности

Для обработки неопределенностей в данной работе применен подход, объединяющий нечеткую логику и метод интервалов с расширяющимися границами.

Если найдена разовая ВК (например, месяц назад), то она представляется как треугольная функция принадлежности с центром в точке ровно месяц назад. Границы функции принадлежности определяются как плюс-минус одна временная единица (день, неделя, месяц или год) от точки на временной линии, к которой это событие относится.

Если найдена продолжительная ВК (например, последние 2 месяца), то она представляется как трапециевидная функция принадлежности. Точкам *b* и *c* на рисунке 5 соответствуют границы нормализованной временной конструкции. Границы функции принадлежности определяются как плюс-

минус одна временная единица (день, неделя, месяц или год) от границы нормализованной временной конструкции. Если точка *c* соответствует настоящему моменту времени, тогда точка *d* также равна настоящему моменту времени.

В каждое правило было добавлено поле 'uncertain'. В данном поле указывается функция, которая определяет интервал неопределенности для ВК, относящейся к данному правилу. Также добавлено поле 'form', где необходимо указать какую форму неопределенности имеет временная конструкция (triangle или trapezoid).

Если перед в тексте ВК содержатся слова 'около', 'примерно', 'приблизительно', 'почти' или 'где-то', то ширина функции принадлежности увеличивается в 2 раза.

Также, когда пациент говорит врачу о том, что какое-то событие произошло с ним 5 лет назад, он более уверен в этом, чем, когда он говорит о том, что какое-то событие произошло с ним 20 лет назад. Во втором случае действительная дата данного события может находиться вне интервала от 19 лет назад до 21 года назад. Для решения данной проблемы был добавлен коэффициент, который увеличивает ширину функции принадлежности для событий, которые произошли более 5 лет назад. Данный коэффициент рассчитывается по следующей формуле:

$$k = 1 + \frac{past - 43800}{175200} \tag{3}$$

Где past – количество прошедшего времени от текущего момента времени в часах;

43800 – количество часов в 5 годах;

175200 – количество часов в 20 годах;

Рисунок 6 демонстрирует распределение событий для 50 случайных пациентов. На оси абсцисс указано количество прошедшего времени от текущего момента времени в часах в логарифмическом масштабе, а на оси ординат – ширина интервала неопределенности в логарифмическом масштабе.

Как можно заметить на графике, большинство событий имеют интервал неопределенности ± месяц и ± год. С увеличением прошедшего времени ширина интервала неопределенности увеличивается благодаря введенному коэффициенту.

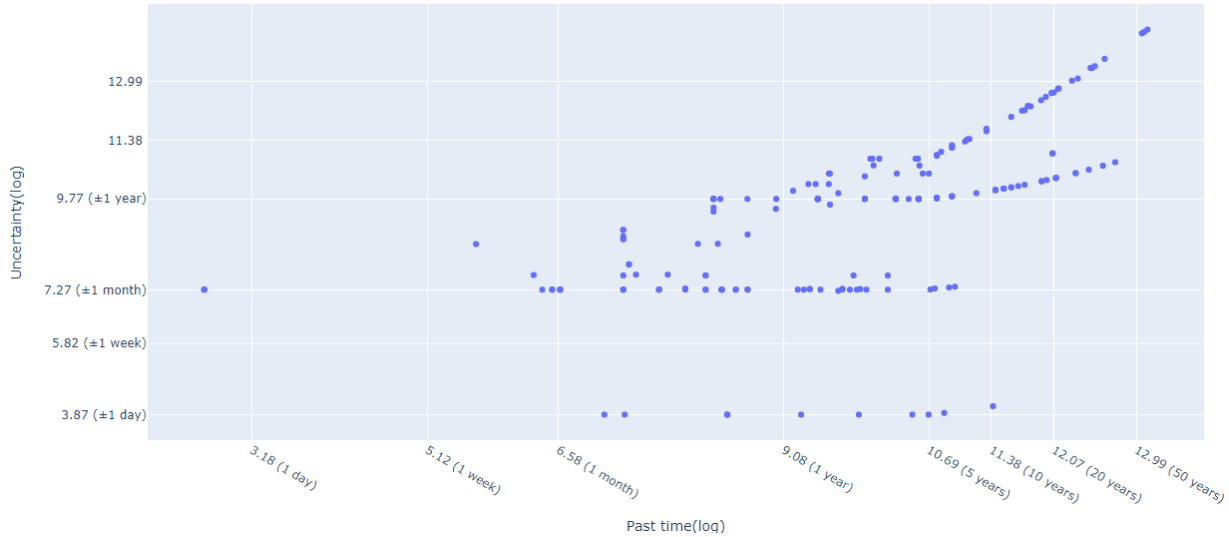


Рисунок 6 – Распределение событий для 50 случайных пациентов

Итоговое правило в разрабатываемом ПО выглядит следующим образом:

```
'r_in_year4d_year_a': {
'pattern': [{"LEMMA": 'в'}, {"TEXT": {"REGEX": yearfull}}, {"TEXT": "г"}],
'norm': lambda ent: strptime('01.07.{}'.format(ent[1].text), '%d.%m.%Y'),
'uncertain': delta_year,
'form': triangle,
'stamp': 1
},
```

Где:

- 'r_halfyear_ago' – уникальное имя правила;
- 'pattern' – Spacy правило для извлечения ВК;
- 'norm' – лямбда функция для нормализации ВК;
- 'uncertain' – функция для расчета интервала неопределенности;
- 'form' – форма неопределенности (triangle или trapezoid);
- 'stamp' – метка.

## 3.5 Тестирование разработанного программного обеспечения

При написании правил для извлечения временных конструкций, создании модуля извлечения событий и оценке точности использовался набор данных НИИ кардиологии г. Томска (7777 предложений). В наборе были отобраны предложения с временными конструкциями.

После написания новых правил, исправления функций нормализации и добавления поиска отрицаний на тестовом наборе данных удалось достичь точности извлечение ВК – 95.5%, нормализации – 94.2%

Для оценки точности была проведен эксперимент на наборе ФГБУ «НМИЦ им. В. А. Алмазова» (2000 предложений), который не использовался при написании правил и создании алгоритма. В отличие от прошлого набора, он содержал предложения с временными конструкциями и без них. На этом наборе данных иногда модуль извлекал фразы, которые не содержат временных конструкций. Для решения данной проблемы был использован обученный классификатор, который определяет включает ли в себя предложение временную конструкцию. После применения данного классификатора было найдено 580 предложений, которые содержат временные конструкции. Затем был применен модуль для извлечения событий. Анализ ошибок модуля приведен в таблице 6. Таким образом, точность извлечения событий на новом наборе данных составила 70%.

Таблица 6 – Анализ ошибок при извлечении событий на наборе данных центра Алмазова

| **Причина ошибки** | **Количество** |
|---|---|
| Ошибка при извлечении события | 77 |
| Ошибка при извлечении временной конструкции | 36 |
| Ошибка синтаксического парсера | 31 |
| Захватил лишние слова | 9 |
| Орфографическая ошибка в предложении | 6 |
| Пропущен предлог | 6 |
| Ошибка парсера событий и извлечения временных конструкций | 3 |
| Грамматическая ошибка в предложении | 2 |
| Сложная грамматическая конструкция | 1 |
| Сложная грамматическая конструкция с грамматическими и орфографическими ошибками в предложении | 1 |
| Сложная грамматическая конструкция с орфографическими ошибками в предложении | 1 |
| Найдены не все однородные члены | 1 |

### 3.6 Результат работы модуля

При написании правил для извлечения временных конструкций, создании модуля. Результат работы модуля можно представить в виде pandas dataframe. Данный формат в последующем можно сохранить в csv или Excel файл. Пример результата работы модуля представлен в таблице 7.

Так как синтаксический парсинг требует большой вычислительной мощности и занимает много времени, имеется возможность сохранить его результаты в виде conll файла. Это позволит при повторной обработке массива данных снизить время работы модуля.

Таблица 7 – Пример результата работы модуля

| Предложение | Дата осмотра | Дата рождения | ВК | Событие | Нормальная форма | Неопределенность | Метка |
|---|---|---|---|---|---|---|---|
| Болеет СД 2 типа в течении 5 лет. | 2010-12-14 18:33:00 | 1957-01-01 00:00:00 | в течении 5 лет | Болеет СД 2 типа | 2005-12-14 18:33:00 – 2010-12-14 18:33:00 | [2004-12-14 18:33:00, 2005-12-14 18:33:00, 2010-12-14 18:33:00, 2010-12-14 18:33:00] | [2] |
| В мае 2009 года в СибГМУ проводилась РЧ - аблация по поводу ЖЭ. | 2010-12-14 21:06:00 | 1950-11-05 00:00:00 | В мае 2009 года | проводилась РЧ - аблация по поводу ЖЭ | 2009-05-15 00:00:00 | [2009-04-15 00:00:00, 2009-05-15 00:00:00, 2009-06-15 00:00:00] | [1] |
| В течение месяца чувствовала себя удовлетворительно - перебоев не было. | 2010-12-14 21:06:00 | 1950-11-05 00:00:00 | В течение месяца | чувствовала себя удовлетворительно | 2010-11-14 21:06:00 – 2010-12-14 21:06:00 | [2010-10-14 21:06:00, 2010-11-14 21:06:00, 2010-12-14 21:06:00, 2010-12-14 21:06:00] | [2] |

Для того, чтобы не обрабатывать каждое предложение в наборе данных. Можно использовать модель машинного обучения, которая определяет есть ли ВК в предложении или нет [30]. Это также поможет снизить время работы модуля при обработке больших массивов данных, но при этом некоторая информация может быть упущена.

## 3.7 Выводы

В ходе разработки программного обеспечения был написан модуль для извлечения временных конструкций и нормализации, который содержит больше 250 правил.

Также был разработан алгоритм для поиска событий, связанных с временными конструкциями, и проанализированы его ошибки. В результате тестирования было обнаружено, что алгоритм не справляется с отрицаниями. Для решения данной проблемы разработан модуль поиска отрицаний.

После этого проведено тестирование программного обеспечения на данных ФГБУ «НМИЦ им. В. А. Алмазова». Точность извлечения событий составила 70%.

# 4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

В данной работе был предложен новый метод для переноса временных конструкции на временную шкалу и предоставления их в удобном для медицинских сотрудников виде. Также разработанные инструменты позволяют собрать необходимые для обучения данные. Алгоритм предполагает использование контекстно-свободных грамматик и современных моделей синтаксического анализа текста с последующей постобработкой выделения ключевых фраз и смыслов из текста.

Потенциальными потребителями разрабатываемого решения являются медицинские учреждения и предприятия, занимающиеся анализом медицинских данных.

Для того, чтобы эффективно использовать научный потенциал проекта необходимо уделять внимание, как разработке, так и проведению её анализа с точки зрения ее востребованности, а также ресурсоэффективности и ресурсосбережения.

Целью данного раздела является проектирование и создание конкурентоспособных разработок, технологий, отвечающих современным требованиям в области ресурсоэффективности и ресурсосбережения.

Достижение цели обеспечивается решением задач:

- оценка коммерческого потенциала и перспективности проведения научных исследований;
- определение возможных альтернатив проведения научных исследований, отвечающих современным требованиям в области ресурсоэффективности и ресурсосбережения;
- планирование научно-исследовательских работ;
- определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования.

## 4.1 Предпроектный анализ

### 4.1.1 Анализ конкурентоспособности технического решения

В виду того, что медицинские информационные системы разрабатываются и используются исключительно в рамках конкретного медицинского учреждения и не распространяются за его пределы, анализ конкурентных технических решений невозможен ввиду отсутствия открытых данных о наличии и свойствах подобных решений. Также существует ряд зарубежных разработок, который представлен в открытом доступе, но данные работы применимы только для медицинских документов на английском языке. Следовательно, данные проекты не применимы в российских медицинских учреждениях.

Анализ конкурентоспособности технического решения был проведен с помощью оценочной карты. В основе данной технологии лежит нахождение средневзвешенного значения показателя качества и перспективности научной разработки К.

Анализ конкурентных технических решений определяется по формуле:

$$К = \sum В_i \cdot Б_i, \quad (4)$$

где К – конкурентоспособность научной разработки или конкурента;

$В_i$ – вес показателя (в долях единицы);

$Б_i$ – балл $i$-го показателя.

Оценочная карта приведена в таблице 8. Из оценочной карты можно сделать вывод о том, что разрабатываемая система является перспективной. Также у разрабатываемого решения наиболее важные критерии имеют высокие показатели, и суммарная оценка составляет 4,12.

Таким образом, данная разработка благодаря своей безопасности, функциональности и повышению производительности труда пользователя может быть конкурентоспособной на рынке.

Таблица 8 – Оценочная карта

| Критерии оценки | Вес критерия | Баллы | Конкурентоспособность |
|---|---|---|---|
| **Технические критерии оценки ресурсоэффективности** | | | |
| 1. Энергоэффективность | 0,05 | 4 | 0,2 |
| 2. Помехоустойчивость | 0,07 | 4 | 0,28 |
| 3. Надежность | 0,15 | 4 | 0,6 |
| 4. Время выполнения алгоритма | 0,05 | 4 | 0,2 |
| 5. Пользовательский интерфейс | 0,05 | 2 | 0,1 |
| 6. Безопасность | 0,1 | 5 | 0,5 |
| 7. Потребность в ресурсах памяти | 0,07 | 4 | 0,28 |
| 8. Функциональная мощность | 0,05 | 5 | 0,25 |
| 9. Простота эксплуатации | 0,08 | 3 | 0,24 |
| 10. Повышение производительности труда пользователя | 0,1 | 5 | 0,5 |
| 11. Качество интеллектуального интерфейса | 0,03 | 4 | 0,12 |
| **Экономические критерии оценки эффективности** | | | |
| 1. Цена | 0,08 | 4 | 0,32 |
| 2. Предполагаемый срок эксплуатации | 0,07 | 4 | 0,28 |
| 3. Послепродажное обслуживание | 0,05 | 4 | 0,25 |
| **Итого** | 1 | 57 | 4,12 |

### 4.1.2 Диаграмма Исикавы

Диаграмма причины-следствия Исикавы позволяет графически проанализировать и сформировать причинно-следственные связи, это инструментальное средство для систематического определения причин проблемы и последующего графического представления.

Проблемной областью анализа является низкая скорость работы врача. На рисунке 7 представлена причинно-следственная диаграмма Исикавы.

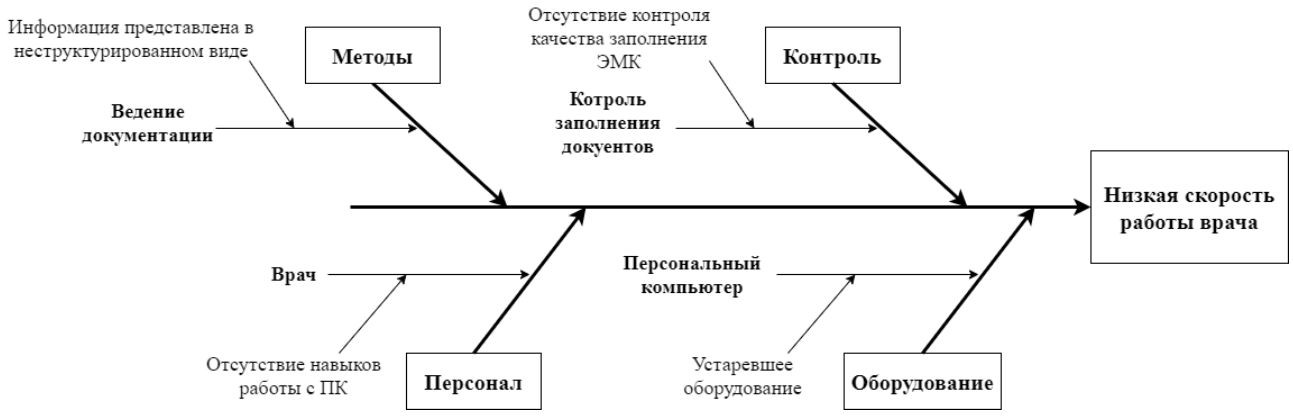


Рисунок 7 – Причинно-следственная диаграмма

Источниками низкой скорости работы врача могут быть персонал, у которого отсутствуют навыки работы с персональным компьютером, устаревшее оборудование, методы работы с медицинской документацией, а также отсутствие контроля качества заполнения документации.

### 4.1.3 SWOT-анализ

Для исследования внутренних и внешних факторов, оценке рисков и конкурентоспособности был проведен комплексный анализ научно-исследовательского проекта – SWOT анализ. Результаты SWOT-анализа приведены в таблице 9.

Необходимо отметить, что особенностью данного решения является его

универсальность и возможность конфигурирования под каждую область медицины. Для решения каждой задачи по отдельности требуется дополнительное конфигурирование, либо предоставление обучающих материалов для специалистов, разрабатывающие конечное решение. Расширение функционала со временем позволит сделать процесс конфигурирования наиболее удобным и эффективным. Поддержание стабильной ценовой политики и оказание услуг по настройки под конкретную предметную область будет напрямую способствовать конкурентоспособности разрабатываемого решения.

Таблица 9 – Итоговая матрица SWOT-анализа

| | **Сильные стороны:**<br>С1. Использование современных моделей синтаксического анализа текста.<br>С2. Возможность построения правил с использованием грамматики зависимостей.<br>С3. Расширяемый функционал.<br>С4. Низкая стоимость. | **Слабые стороны:**<br>Сл1. Для работы с решением необходим специалист с опытом в области обработки естественных языков.<br>Сл2. Данная библиотека не является конечным решением актуальных задач, а лишь конструктором этого решения.<br>Сл3. Недостаток положительных отзывов. |
|---|---|---|
| **Возможности:**<br>В1. Потребность рынка решении задач по извлечению информации<br>В2. Наличие других языков, в которых синтаксические модели показывают высокое качество | – Добавление поддержки новых языков позволит выйти;<br>на зарубежный рынок<br>– Добавления новых конструкций для правил с учетом потребностей пользователей сделает конфигурирование более гибким;<br>– Привлечение новых клиентов благодаря конфигурированию продукта под их задачу. | – Разработка различных обучающих материалов: подробная документации с примерами, разработка курсов, проведение семинаров;<br>– Привлечение специалистов, внешних консультантов, а также профессиональное сообщество для масштабирования с точки зрения функционала и других языков. |
| **Угрозы:**<br>У1. Появление на рынке новых игроков с аналогичной разработкой.<br>У2. Появление новых, принципиально отличающихся методов решения. | – Регулярное исследование новых решений и моделей с последующим внедрением в решение;<br>– Работа над оптимизацией производительности<br>решения. | – Оказание услуг по конфигурации под конкретную задачу;<br>– Поддержание стабильной ценовой политики. |

## 4.2 Инициация проекта

В данном разделе приводится информация о заинтересованных сторонах научно-исследовательского проекта, иерархии целей проекта и критериях достижения целей.

### 4.2.1 Цели и результаты проекта

В процессе инициации проекта определяются начальные цели и определяются внутренние и внешние заинтересованные стороны проекта,

которые будут взаимодействовать и влиять на общий результат научного проекта. Заинтересованные стороны проекта – это лица или организации, которые активно участвуют в проекте или интересы которых могут быть затронуты как положительно, так и отрицательно в ходе исполнения или в результате завершения проекта. Информация по заинтересованным сторонам научно-исследовательского проекта представлена в таблице 10.

Таблица 10 – Заинтересованные стороны научно-исследовательского проекта

| Заинтересованные стороны научно-исследовательского проекта | Ожидания заинтересованных сторон |
|---|---|
| Отделение информационных технологий ТПУ | Научные публикации. Защита магистерской диссертации. |
| Медицинский персонал | Уменьшение времени работы с медицинской документацией, благодаря автоматизации процесса извлечения временных конструкций из электронных медицинских карт. |
| Пациенты | Получение более качественных консультаций врачей, благодаря уменьшению работы врача с документацией. |
| Научное сообщество | Алгоритм извлечения временных конструкций из электронных медицинских карт. Инструмент для сбора данных для моделей машинного обучения. |

В таблице 11 представлена информация об иерархии целей научно-исследовательского проекта и критериях их достижения.

Таблица 11 – Иерархия целей научно-исследовательского проекта и критерии их достижения

| | |
|---|---|
| Цели проекта: | Разработать алгоритмы и программное обеспечение идентификации временных конструкций в слабоструктурированных электронных медицинских текстах. |
| Ожидаемые результаты проекта: | Алгоритмы и программное обеспечение идентификации временных конструкций в слабоструктурированных электронных медицинских текстах. |
| Требования к результату проекта: | Данные подготовлены для анализа. |
| | Алгоритм извлечения временных конструкций из электронных медицинских карт. |
| | Бесперебойная работа программных модулей проекта. |
| | Формализованное описание работы программных модулей проекта. |

### 4.2.2 Организационная структура проекта

На данном этапе работы определяется состав рабочей группы научно-исследовательского проекта, определяются роли каждого участника в

данном проекте. В таблице 12 определены участники научно-исследовательского проекта и их роли.

Таблица 12 – Рабочая группа проекта

| № п/п | ФИО, основное место работы, должность | Роль в проекте | Функции |
|---|---|---|---|
| 1 | Аксёнов Сергей Владимирович, ОИТ ИШИТР ТПУ, доцент | Научный руководитель | Составление научных задач, контроль выполнения проекта, проверка документации |
| 2 | Журман Дмитрий Александрович | Инженер | Проектирование, реализация |

## 4.3 Планирование управления научно-техническим проектом

### 4.3.1 Структура работы в рамках проекта

Для составления структуры работ определяются ключевые события проекта, затем детальный перечень этапов и работ. На каждый вид работ определяется исполнитель. Распределение исполнителей по данным видам работ приведено в таблице 13.

Таблица 13 – Перечень этапов, работ и распределение исполнителей

| Основные этапы | № | Содержание работы | Должность исполнителя |
|---|---|---|---|
| Разработка задания | 1 | Постановка задачи | Руководитель, Инженер |
| Теоретические исследования | 2 | Подбор и изучение материалов по теме | Инженер |
| | 3 | Разработка и утверждения ТЗ | Руководитель, Инженер |
| | 4 | Календарное планирование работ по теме | Инженер |
| | 5 | Разработка вариантов исполнения проекта | Руководитель, Инженер |
| Разработка и проектирование системы | 6 | Подготовка данных | Инженер |
| | 7 | Подбор модели синтаксического парсера для русского языка | Инженер |
| | 8 | Разработка модуля для извлечения временных конструкций | Инженер |
| | 9 | Разработка модуля для поиска событий, связанных с временными конструкциями | Инженер |
| | 10 | Проверка работы библиотеки на тестовых данных и правилах | Руководитель, Инженер |
| Оформление документации | 11 | Составление документации | Инженер |

### 4.3.2 Структура работы в рамках проекта

Трудовые затраты в большинстве случаях образуют основную часть стоимости разработки, поэтому важным моментом является определение трудоемкости работ каждого из участников научного исследования.

Трудоемкость выполнения научного исследования оценивается экспертным путем в человеко-днях и носит вероятностный характер, т.к. зависит от множества трудно учитываемых факторов. Для определения ожидаемого (среднего) значения трудоемкости $t_{ожi}$ используется следующая формула:

$$t_{ожi} = \frac{3t_{min i} + 2t_{max i}}{5}, \quad (5)$$

где $t_{ожi}$ – ожидаемая трудоемкость выполнения i-ой работы чел.-дн.;

$t_{min i}$ – минимально возможная трудоемкость выполнения заданной i-ой работы (оптимистическая оценка: в предположении наиболее благоприятного стечения обстоятельств), чел.-дн.;

$t_{max i}$ – максимально возможная трудоемкость выполнения заданной i-ой работы (пессимистическая оценка: в предположении наиболее неблагоприятного стечения обстоятельств), чел.-дн.

Исходя из ожидаемой трудоемкости работ, определяется продолжительность каждой работы в рабочих днях $T_р$, учитывающая параллельность выполнения работ несколькими исполнителями. Такое вычисление необходимо для обоснованного расчета заработной платы, так как удельный вес зарплаты в общей сметной стоимости научных исследований составляет около 65 %.

$$T_{р_i} = \frac{t_{ожi}}{Ч_i}, \quad (6)$$

где $T_{рi}$ – продолжительность одной работы, раб. дн.;

$t_{ожi}$ – ожидаемая трудоемкость выполнения одной работы, чел.-дн.

$Ч_i$ – численность исполнителей, выполняющих одновременно одну и ту же работу на данном этапе, чел.

Наиболее удобной и наглядной формой представления графика проведения работ является ленточный график в форме диаграммы Ганта. Для удобства построения графика, длительность каждого из этапов работ из рабочих дней следует перевести в календарные дни. Для этого необходимо воспользоваться формулой:

$$T_{кi} = T_{рi} \cdot k_{кал}, \quad (7)$$

где Tкi – продолжительность выполнения i-ой работы в календарных днях;

Tpi – продолжительность выполнения i-ой работы в рабочих днях;

$k_{кал}$ – коэффициент календарности.

Коэффициент календарности определяется по формуле:

$$k_{кал} = \frac{T_{кал}}{T_{кал} - T_{вых} - T_{пр}}, \quad (8)$$

где $T_{кал}$ – количество календарных дней в году;

$T_{вых}$ – количество выходных дней в году;

$T_{пр}$ – количество праздничных дней в году.

В соответствии с производственным календарем (для 6-дневной рабочей недели) в 2019 году 365 календарных дней, 299 рабочих дней, 66 выходных/праздничных дней. Полученные данные сведены в таблицу 14.

Таблица 14 – Временные показатели проведения научного исследования

| Название работы | Трудоемкость работы | | | | | | Длительность работы в рабочих днях, $T_{pi}$ | | Длительность работы в календарных днях, $T_{ki}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $t_{min}$, чел-дни | | $t_{max}$,чел-дни | | $t_{ож}$, чел-дни | | | | | |
| | Руководитель | Инженер | Руководитель | Инженер | Руководитель | Инженер | Руководитель | Инженер | Руководитель | Инженер |
| Постановка задачи | 1 | 2 | 3 | 4 | 1,8 | 3,2 | 2 | 3 | 2 | 4 |
| Подбор и изучение материалов по теме | | 7 | | 9 | | 8,2 | | 8 | | 10 |
| Разработка и утверждения ТЗ | 1 | 9 | 3 | 11 | 1,8 | 9,8 | 2 | 10 | 2 | 12 |
| Календарное планирование работ по теме | | 1 | | 3 | | 1,8 | | 2 | | 2 |
| Разработка вариантов исполнения проекта | 3 | 5 | 5 | 7 | 4,2 | 6,2 | 4 | 6 | 5 | 8 |
| Подготовка данных | | 11 | | 13 | | 12,2 | | 12 | | 15 |
| Подбор модели синтаксического парсера для русского языка | | 11 | | 13 | | 12,2 | | 12 | | 15 |
| Разработка модуля для извлечения временных конструкций | | 11 | | 13 | | 12,2 | | 12 | | 15 |
| Разработка модуля для поиска событий, связанных с временными конструкциями | | 11 | | 13 | | 12,2 | | 12 | | 15 |
| Проверка работы библиотеки на тестовых данных и правилах | 1 | 5 | 3 | 7 | 1,8 | 6,2 | 2 | 6 | 2 | 8 |
| Составление документации | | 6 | | 8 | | 7,2 | | 7 | | 9 |

### 4.3.3 Разработка графика проведения научного исследования

На основе полученной таблицы строится календарный план-график. График строится для максимального по длительности исполнения работ с разбивкой по месяцам и декадам. График работ приведен в приложении Б.

## 4.4 Бюджет научного и исследования

Для полноты и достоверности учета всех расходов сгруппируем все затраты по следующим статьям

- затраты на материалы;
- затраты на амортизацию;
- основная заработная плата исполнителей;
- дополнительная заработная плата исполнителей темы;
- отчисления во внебюджетные фонды (страховые отчисления);
- накладные расходы.

### 4.4.1 Расчет материальных затрат

Расчет материальных затрат осуществляется по следующей формуле:

$$З_м = (1 + k_T) \cdot \sum_{i=1}^{m} Ц_i \cdot N_{расхi} , \quad (9)$$

где m – количество видов материальных ресурсов, потребляемых при выполнении научного исследования;

$N_{расхi}$ – количество материальных ресурсов i-го вида;

$Ц_i$ – цена приобретения единицы i-го вида потребляемых материальных ресурсов;

$k_T$ – коэффициент, учитывающий транспортно-заготовительные расходы.

Материальные затраты приведены в таблице 15.

Таблица 15 – Материальные затраты

| Наименование | Единица измерения | Количество | Цена за ед.,руб. | Сумма, руб. |
|---|---|---|---|---|
| Ручка шариковая | шт. | 2 | 12 | 24 |
| Тетрадь в клетку 24 листа | шт. | 1 | 24 | 24 |
| Бумага А4, 500 листов | шт. | 1 | 321 | 321 |
| Картридж TN1075 | шт. | 1 | 410 | 410 |
| Всего за материалы | | | | 779 |
| Транспортно-заготовительные расходы (3%) | | | | 23,37 |
| Итого | | | | 802,37 |

### 4.4.2 Расчет затрат на специальное оборудование для научных (экспериментальных) работ

В данную статью включают все затраты, связанные с приобретением специального оборудования (приборов, контрольно-измерительной техники, стендов, устройств и механизмов), однако оборудование специально для проекта не приобреталось, поэтому была рассчитана амортизация оборудования на время проекта.

Для работы над проектом использовался ноутбук. Амортизацию рассчитаем линейным способом. Первоначальная стоимость ПК 182500 рублей; срок полезного использования для машин офисных код 330.28.23.23 составляет 2-3 года, берем 3 года; планируется использовать ПК для написания работы в течение 4 месяцев. Тогда месячная норма амортизации:

$$Ан = \frac{1}{n} * 100\% = \frac{1}{12*3} * 100\% = 2,8, \quad (10)$$

где $n$ - количество месяцев полезного срока эксплуатации ОС.

Ежемесячные амортизационные отчисления:

$$\text{Ам} = 182500 * 2,8\% = 5110 \text{ рублей}$$

Итоговая сумма амортизации основных средств:

$$\text{А} = 5110 * 4 = 20440 \text{ рублей}$$

Таким образом, сумму амортизации основных средств составила 20440 рублей.

### 4.4.3 Основная заработная плата

В настоящую статью включается основная заработная плата научных и инженерно-технических работников, рабочих макетных мастерских и опытных производств, непосредственно участвующих в выполнении работ по данной теме. Величина расходов по заработной плате определяется исходя из трудоемкости выполняемых работ и действующей системы окладов и

тарифных ставок. В состав основной заработной платы включается премия, выплачиваемая ежемесячно из фонда заработной платы в размере 20 –30 % от тарифа или оклада.

Статья включает основную заработную плату работников, непосредственно занятых выполнением НТИ, и дополнительную заработную плату:

$$З_{ЗП} = З_{осн} + З_{доп}, \quad (11)$$

где $З_{осн}$ – основная заработная плата;

$З_{доп}$ – дополнительная заработная плата (12-20% от основной).

Основная заработная плата руководителя (лаборанта, инженера) от предприятия рассчитывается по следующей формуле:

$$З_{осн} = З_{дн} \cdot T_{р}, \quad (12)$$

где $З_{осн}$ – основная заработная плата одного работника;

$З_{дн}$ – среднедневная заработная плата работника, руб.;

$Т_р$ – продолжительность работ, выполняемых работником, раб.дн.

Среднедневная зарплата рассчитывается по формуле:

$$З_{дн} = \frac{З_{м} \cdot М}{F_{д}}, \quad (13)$$

где $З_м$ – месячный должностной оклад работника, руб.;

М – количество месяцев работы без отпуска в течение года:

при отпуске в 48 раб. дней М = 10,4 месяца, 6-дневная неделя;

$F_д$ – действительный годовой фонд рабочего времени научно-технического персонала, раб. дн.

Баланс рабочего времени приведен в таблице 16.

Таблица 16 – Баланс рабочего времени

| **Показатели рабочего времени** | **Руководитель** | **Студент** |
|---|---|---|
| Календарное число дней | 365 | 365 |
| Количество нерабочих дней<br>-выходные дни<br>-праздничные дни | <br>52<br>14 | 52<br>14 |
| Потери рабочего времени (отпуск/невыходы по болезни) | 48 | 48 |
| Действительный годовой фонд рабочего времени | 251 | 251 |

Месячный должностной оклад работника:

$$З_М = З_{ТС} \cdot (1 + k_{пр} + k_д) \cdot k_р, \quad (14)$$

где $З_{тс}$ – заработная плата по тарифной ставке, руб.;

$k_{пр}$ – премиальный коэффициент, равный 0,3;

$k_д$ – коэффициент доплат и надбавок составляет примерно 0,2-0,5;

$k_р$ – районный коэффициент, равный 1,3 для Томска.

Расчет основной платы представлен в таблице 17.

Таблица 17 – Расчет основной заработной платы

| **Исполнители** | **Оклад** | $k_{пр}$ | $k_д$ | $k_р$ | **$З_м$, руб** | **$З_{дн}$, руб.** | **$Т_р$, раб. дн.** | **$З_{осн}$, руб.** |
|---|---|---|---|---|---|---|---|---|
| Руководитель | 33664 | 0,3 | 0,4 | 1,3 | 74397,44 | 3082,60 | 10 | 30826,03 |
| Инженер | 21760 | 0,3 | 0,4 | 1,3 | 48089,6 | 1992,56 | 90 | 179330,14 |
| Итого: | | | | | | | | 210156,17 |

### 4.4.4 Дополнительная заработная плата исполнителей темы

Расчет дополнительной заработной платы ведется по формуле:

$$З_{доп} = k_{доп} \cdot З_{осн}, \quad (15)$$

где $k_{доп}$ – коэффициент дополнительной заработной платы, на стадии проектирования принимается равным 0,12. Расчеты дополнительной заработной представлены в таблице 18.

Таблица 18 – Расчет дополнительной заработной платы

| **Заработная плата** | **Руководитель** | **Инженер** |
|---|---|---|
| Основная зарплата | 30826,03 | 179330,14 |
| Дополнительная зарплата | 3699,12 | 21519,62 |
| Зарплата исполнителя | 34525,15 | 200849,76 |
| Итого: | | 235374,91 |

### 4.4.5 Отчисление во внебюджетные фонды

В данной статье отражаются обязательные отчисления по установленным законодательствам Российской Федерации нормам органам государственного социального страхования, пенсионного фонда и медицинского страхования.

Величина отчислений во внебюджетные фонды определяется исходя из

следующей формулы:

$$З_{внеб} = k_{внеб} \cdot (З_{осн} + З_{доп}) , \quad (16)$$

где $k_{внеб}$ - коэффициент отчислений на уплату во внебюджетные фонды.

Расчет отчислений приведен в таблице 19.

Таблица 19 – Расчет отчисления во внебюджетные фонды

| Исполнитель | Зарплата исполнителя, руб. | Коэффициент отчислений | Отчисления |
|---|---|---|---|
| Руководитель | 34525,15 | 0,302 | 10426,60 |
| Инженер | 200849,76 | | 60656,63 |
| Итого | | | 71083,23 |

### 4.4.6 Накладные расходы

В накладные расходы должны быть включены те затраты организации, которые не попали в предыдущие статьи расходов: оплата электроэнергии, услуг связи, размножение материалов, печать и ксерокопирование материалов и т.д.

Накладные расходы определяются по следующей формуле:

$$З_{накл} = k_{нр} \cdot (З_{осн} + З_{доп}), \quad (17)$$

где $k_{нр}$ – коэффициент, учитывающий накладные расходы, 80 %.

Таким образом, накладные расходы составили 188299,98 рублей.

### 4.4.7 Прямые затраты

В этих расходах нужно посчитать затраты на электроэнергию, потребляемую оборудованием. Для этого нужно узнать мощность, время использования оборудования и рассчитать затраты.

Стоимость использования сети Internet – 500 рублей в месяц.

Время использования сети Internet – 4 месяца.

Расходы на использование сети Internet – 2000 рублей.

Стоимость 1 кВт – составляет 3,5 руб.

Средняя мощность, потребляемая электрооборудованием во время работы, – 750 Вт.

Время использования электрооборудования составляет 720 часов.

Расходы на использование электроэнергии – 1890 рублей.

Таким образом, прямые расходы составили 3890 рублей.

### 4.4.8 Формирование бюджета затрат научно-исследовательского проекта

Рассчитанная величина затрат научно-исследовательской работы является основой для формирования бюджета затрат проекта, который при формировании договора с заказчиком защищается научной организацией в качестве нижнего предела затрат на разработку научно-технической продукции.

Определение бюджета затрат приведено в таблице 20.

Таблица 20 – Бюджет затрат

| Наименование статьи | Сумма, руб. | Удельный вес, % |
|---|---|---|
| 1. Материальные затраты НТИ | 802,37 | 0,15 |
| 2.Затраты на специальное оборудование для научных работ | 20440 | 3,93 |
| 3. Затраты по основной заработной плате исполнителей темы | 210156,17 | 40,42 |
| 4.Затраты по дополнительной заработной плате исполнителей темы | 25215,74 | 4,85 |
| 5.Отчисление во внебюджетные фонды | 71083,23 | 13,67 |
| 6. Накладные расходы | 188299,98 | 36,22 |
| 7. Прямые расходы | 3890 | 0,75 |
| 8. Бюджет затрат НТИ | 519887,49 | 100,00 |

## 4.5 Риски

Риски в реализации проекта включают в себя возможные неопределенные события, которые могут возникнуть в проекте и вызвать последствия, которые повлекут за собой нежелательные эффекты. Оценка рисков проекта представлена в приложении В. Для каждого из них даны рекомендации по смягчению их воздействия.

В результате данного этапа были рассмотрены возможные риски при реализации настоящей работы. Основная часть рисков может привести к невостребованности разработанного решения. Однако их воздействие можно минимизировать благодаря проведению прототипирования,

итеративности разработки, проведению технического анализа стоимости и проведению сравнительного тестирования.

## 4.6 Выводы

В результате проведения исследования по разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение» были определены показатели затрат научно-исследовательской работы.

В данной главе был разработан план и сформирован бюджет технического решения. Продолжительность проекта составила 90 рабочих дня, а общий бюджет затрат составил 519 887 рублей. Основная часть затрат приходится на заработную плату исполнителей проекта. Таким образом план-график и бюджет проекта успешно укладываются в ограничения заказчика. Разработанный реестр рисков отражает потенциальные пути преодоления внешних и внутренних рисков и способствует успешной реализации проекта, а также его дальнейшее существование.

# 5. Социальная ответственность

Объектом исследования и разработки является программная библиотека для извлечения временных конструкций из электронных медицинских документов. Полученная библиотека позволит перенести данные конструкции на временную шкалу и представить их в удобном для медицинских сотрудников виде. Также разработанные инструменты позволяют собрать необходимые для обучения данные. Теоретическая значимость данной работы заключается в разработке алгоритмов и методов по извлечению значимой информации из электронных медицинских карт, не прибегая к методам машинного обучения.

Потенциальными потребителями разрабатываемого решения являются предприятия, занимающиеся анализом медицинских данных.

Разработка данной библиотеки осуществлялась в лабораторных условиях, так как для работы алгоритма необходим высокопроизводительный графический процессор.

## 5.1 Правовые и организационные вопросы обеспечения безопасности

### 5.1.1 Специальные (характерные для проектируемой рабочей зоны) правовые нормы трудового законодательства

Согласно СП 2.4.3648-20 “Санитарно-эпидемиологические требования к организациям воспитания и обучения, отдыха и оздоровления детей и молодежи” с 2021 г. вопрос установления перерывов во время работы за компьютером нормативно не урегулирован [31]. Работодатель может самостоятельно установить порядок предоставления перерывов в работе за компьютером для отдыха в правилах внутреннего трудового распорядка. Указанные перерывы включаются в рабочее время. То есть они не продлевают продолжительность рабочего дня сотрудника. Согласно трудовому кодексу

Российской Федерации от 30.12. 2001 г. № 197–ФЗ (ред. от 01.04.2019 г.) во время этих перерывов работник не должен выполнять другую работу [32].

Так как работа с данной библиотекой на предприятии подразумевает сбор и анализ персональных данных. Чтобы ограничить доступ к медицинским данным и обеспечить их безопасность, обработка данных должна осуществляться в соответствии с федеральным законом о защите персональных данных [33]:

1. Обработка персональных данных должна осуществляться на законной и справедливой основе.

2. Обработка персональных данных должна ограничиваться достижением конкретных, заранее определенных и законных целей. Не допускается обработка персональных данных, несовместимая с целями сбора персональных данных.

3. Не допускается объединение баз данных, содержащих персональные данные, обработка которых осуществляется в целях, несовместимых между собой.

4. Обработке подлежат только персональные данные, которые отвечают целям их обработки.

5. Содержание и объем обрабатываемых персональных данных должны соответствовать заявленным целям обработки. Обрабатываемые персональные данные не должны быть избыточными по отношению к заявленным целям их обработки.

### 5.1.2 Организационные мероприятия при компоновке рабочей зоны

Работа с разрабатываемой библиотекой подразумевает, что оператор осуществляет взаимодействие с ней сидя за ПК.

Конструкция рабочего места и взаимное расположение всех его элементов (сиденье, стол, средства отображения информации и т.д.) должны

соответствовать общим эргономическим требованиям, приведённым в таблице 21 [34].

Таблица 21 – Нормы оборудования рабочих мест

| | |
|---|---|
| Ширина рабочего стола | не менее 600 мм |
| Глубина рабочего стола | не менее 1200 мм |
| Высота рабочего стола | От 680 до 800 мм (если высота стола не регулируется – 725 мм) |
| Угол наклона спинки | в пределах 0°±30° от вертикального положения |
| Расстояние спинки от переднего края сиденья | от 260 до 400 мм |
| Высота поверхности сиденья | от 400 до 550 мм |
| Сидение | Ширина и глубина не менее 400 мм |
| Подставка для ног | Ширина – от 300 мм, глубина – от 400 мм, с углом наклона до 20 градусов |
| Расстояние клавиатуры от края стола | от 100 до 300 мм |

Согласно ГОСТ Р 50923-96 “Дисплеи. Рабочее место оператора” дисплей на рабочем месте оператора должен располагаться так, чтобы изображение в любой его части было различимо без необходимости поднять или опустить голову. Дисплей на рабочем месте должен быть установлен ниже уровня глаз оператора. Угол наблюдения экрана оператором относительно горизонтальной линии взгляда не должен превышать 60°.

Клавиатура на рабочем месте оператора должна располагаться так, чтобы обеспечивалась оптимальная видимость экрана. Также клавиатура должна иметь возможность свободного перемещения [34].

## 5.2 Производственная безопасность

В подразделе проанализированы вредные и опасные факторы, которые могут возникать при проведении исследований в лаборатории, при разработке или эксплуатации проектируемого решения.

### 5.2.1 Анализ вредных и опасных факторов, которые может создать объект исследования

Перечень опасных и вредных факторов, характерных для объекта исследования представлен в таблицах 22 и 23 согласно ГОСТ 12.0.003-2015

“Система стандартов безопасности труда (ССБТ). Опасные и вредные производственные факторы. Классификация.” [35].

Таблица 22 – Возможные вредные факторы

| Факторы (ГОСТ 12.0.003-2015) | Этапы работ | | | Нормативные документы |
|---|---|---|---|---|
| | Разработка | Изготовление | Эксплуатация | |
| 1.Перенапряжение анализаторов, в том числе вызванное информационной нагрузкой | + | + | + | 1. Трудовой кодекс Российской Федерации от 30.12. 2001 г. № 197–ФЗ (ред. от 01.04.2019 г.).<br>2. СП 52.13330.2016. Естественное и искусственное освещение.<br>3. СанПиН 1.2.3685-21 "Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания"<br>4. ГОСТ Р 50923-96 Дисплеи. Рабочее место оператора.<br>5. ГОСТ Р 50948-2001. Средства отображения информации индивидуального пользования. Общие эргономические требования и требования безопасности.<br>6. ГОСТ Р ИСО 9241-5-2009. Эргономические требования к проведению офисных работ с использованием видеодисплейных терминалов. |
| 2.Отсутствие или недостаток естественного света | + | + | + | |
| 3.Недостаточная освещенность рабочей зоны | + | + | + | |
| 4. Превышение уровня шума на рабочем месте. | + | + | + | |
| 5. Повышенный уровень электромагнитных излучений. | + | + | + | |
| 6.Статические перегрузки, связанные с рабочей позой | + | + | + | |

Таблица 23 – Возможные опасные факторы

| Факторы (ГОСТ 12.0.003-2015) | Этапы работ | | | Нормативные документы |
|---|---|---|---|---|
| | Разработка | Изготовление | Эксплуатация | |
| 1. Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека | + | + | + | ГОСТ 12.1.038-82 ССБТ. Предельно допустимые значения напряжений прикосновения и токов |

### 5.2.2 Перенапряжение анализаторов, в том числе вызванное информационной нагрузкой

Основная характеристика анализаторов – высокая чувствительность, хотя не всякий раздражитель, действующий на анализатор, вызывает ощущение. Чтобы ощущение проявилось, необходима определенная интенсивность раздражителя. Всякое воздействие, превышающее предел интенсивности, вызывает боль и нарушение деятельности анализаторов.

1) Источник возникновения фактора – поступающая информация с монитора компьютера;

2) воздействие фактора на организм человека – интенсивное или длительное воздействие перенапряжение анализаторов может привести к функциональному чрезмерному напряжению, стать причиной профессиональных заболеваний;

3) допустимые нормы – с 2021 года вопрос установления перерывов во время работы за компьютерами нормативно не урегулирован. Работодатель может самостоятельно установить порядок предоставления перерывов в работе за компьютером для отдыха в правилах внутреннего трудового распорядка. Указанные перерывы включаются в рабочее время. То есть они не продлевают продолжительность рабочего дня сотрудника. Во время этих перерывов работник не должен выполнять другую работу. Перерыв предоставляется ему для отдыха. Также перерывы в работе для отдыха от компьютера нужно предоставлять отдельно от перерыва на обед согласно трудовому кодексу Российской Федерации [32].

4) предлагаемыми средствами защиты для минимизации воздействия фактора являются регулярные перерывы для сотрудников, работающих с данной библиотекой.

### 5.2.3 Отсутствие или недостаток естественного света и недостаточная освещенность рабочей зоны.

Помещения должны иметь как естественное, так и искусственное освещение. Согласно СП 52.13330.2016 “Естественное и искусственное освещение” естественное освещение осуществляется через светопроемы, обеспечивающие необходимый коэффициент естественной освещенности (КЕО) не ниже 1,2 % [36].

1) Источник возникновения фактора – вредное воздействие параметров освещения проявляется в отсутствии или недостатке естественного света, а также недостаточной освещенности рабочей зоны;

2) воздействие фактора на организм человека – недостаточное освещение влияет на функционирование зрительного аппарата, то есть определяет зрительную работоспособность, на психику человека, его эмоциональное состояние, вызывает усталость центральной нервной системы, возникающей в результате прилагаемых усилий для опознания четких или сомнительных сигналов;

3) допустимые нормы: согласно СанПиН 1.2.3685-21 "Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания" освещенность рабочего места в кабинетах, рабочих комнатах и офисах при работе за ЭВМ в горизонтальной плоскости от общего искусственного освещения должна быть 300 [37];

4) предлагаемые средства защиты – к средствам нормализации освещенности рабочих мест относятся: источники света, осветительные приборы, световые проемы, светозащитные устройства, светофильтры, защитные очки.

Ниже приведены расчеты для создания освещенности Е = 300 ЛК для лабораторного помещения, в котором изготовлялась библиотека.

Размеры помещения: А = 4,8 м, ширина В = 4,3 м, высота $H$ = 2,8 м. Высота рабочей поверхности $h_{рп}$ = 0,8 м.

Коэффициент отражения стен $R_с$ = 30 %, потолка $R_п$ = 50 %. Коэффициент запаса k = 1,5, коэффициент неравномерности Z = 1,1. Рассчитываем систему общего люминесцентного освещения.

Для заданной высоты помещения подойдут двухламповые светильники ШОД с λ = 1,2.

Приняв $h_c$ = 0,5 м, определяем расчетную высоту:

$$h = H - h_c - h_{\text{рп}} = 2{,}8 - 0{,}5 - 0{,}8 = 1{,}5 \text{ м}; \quad (18)$$

Расстояние между светильниками:

$$L = \lambda \cdot h = 1{,}5 \cdot 1{,}2 = 1{,}8 \text{ м}; \quad (19)$$

Расстояние от крайнего ряда светильников до стены:

$$\frac{L}{3} = 0{,}6 \text{ м}. \quad (20)$$

Определяем количество рядов светильников и количество светильников в ряду:

$$n_{\text{ряд}} = \frac{(B - \frac{2}{3}L)}{L} + 1 = \frac{(4{,}3 - \frac{2}{3} \cdot 1{,}8)}{1.8} + 1 \approx 3. \quad (21)$$

$$n_{\text{св}} = \frac{(A - \frac{2}{3}L)}{l_{\text{св}} + 0{,}5} = \frac{(4{,}8 - \frac{2}{3} \cdot 1{,}8)}{1{,}228 + 0{,}5} \approx 2. \quad (22)$$

Размещаем светильники в три ряда. В каждом ряду можно установить 2 светильника типа ШОД мощностью 40 Вт (с длиной 1,228 м), при этом разрывы между светильниками в ряду составят 1,144 м. Изображаем в масштабе план помещения и размещения на нем светильников (рисунок 8). Учитывая, что в каждом светильнике установлено две лампы, общее число ламп в помещении N = 12.

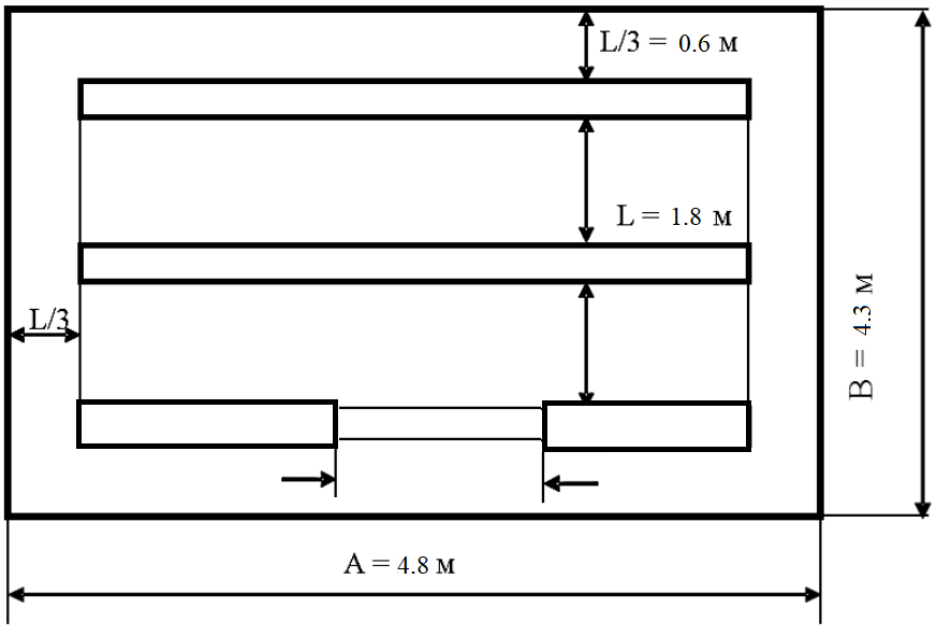


Рисунок 8. План помещения и размещения светильников с люминесцентными лампами

Находим индекс помещения:

$$i = \frac{S}{h(A+B)} = \frac{20{,}64}{2{,}8(4{,}8+4{,}3)} = 0{,}81 \quad (23)$$

Определяем коэффициент использования светового потока: η = 0,3.

Определяем потребный световой поток ламп в каждом из рядов:

$$\Phi = \frac{E_н \cdot S \cdot K_з \cdot Z}{N_л \cdot \eta} = \frac{300 \cdot 20{,}64 \cdot 1{,}5 \cdot 1{,}1}{12 \cdot 0{,}3} = 2838 \quad (24)$$

Выбираем ближайшую стандартную лампу – ЛТБ 40 Вт с потоком 2850 лм [38]. Делаем проверку выполнения условия:

$$-10\% \leq \frac{\Phi_{л.станд} - \Phi_{л.расч}}{\Phi_{л.станд}} \cdot 100\% \leq 20\% \quad (25)$$

Получаем: –10 % ≤ 0,42 % ≤ +20 %.

Определяем электрическую мощность осветительной установки

$$P = 12 \cdot 40 = 480 \text{ Вт}. \quad (26)$$

Таким образом, для разрабатываемого помещения необходимо 6 светильников типа ШОД мощностью 40 Вт. Учитывая, что в каждом светильнике установлено две лампы, в помещении требуется установить 12 ламп ЛТБ 40 Вт с потоком 2850 лм.

### 5.2.4 Статические перегрузки, связанные с рабочей позой

1) Источник возникновения фактора – рабочее место;

2) воздействие фактора на организм человека – неправильная рабочая поза может привести к хроническому спазму (повышенной напряженности) мышц руки, невралгии, плекситу, обострению шейного и грудного радикулита и ряду других неврологических заболеваний;

3) допустимые нормы согласно ГОСТ Р ИСО 9241-5-2009 “Эргономические требования к проведению офисных работ с использованием видеодисплейных терминалов”:

- бедра расположены приблизительно в горизонтальной позиции, а ноги от колена до ступни - в вертикальной позиции; высота сиденья должна равняться длине голеней пользователя до подколенной области или быть немного меньше;
- плечо расположено вертикально, предплечье – горизонтально;
- работа не требует сгибаний или разгибаний запястий;
- позвоночник расположен вертикально;
- ступня составляет угол в 90° по отношению к подколенной части ноги;
- скручивание верхней части туловища отсутствует;
- линия зрения заключена между горизонталью и 60° ниже горизонтали [39].

4) предлагаемые средства защиты – правильная организация рабочего места.

### 5.2.5 Превышение уровня шума на рабочем месте

1) Источник возникновения фактора – вентиляционные установки, кондиционеры, ЭВМ и его периферийные устройства, а также серверные комнаты;

2) воздействие фактора на организм человека – оказывает раздражающее влияние на работника, повышает его утомляемость, а при выполнении задач, требующих внимания и сосредоточенности, способен

привести к росту ошибок и увеличению продолжительности выполнения задания;

3) допустимые нормы: согласно нормам, указанным в ГОСТ Р 50923-96 “Дисплеи. Рабочее место оператора” нормативным эквивалентным уровнем звука на рабочих местах является 50 дБА [34].

4) предлагаемые средства защиты – снизить уровень шума в помещениях можно использованием звукопоглощающих материалов с максимальными коэффициентами звукопоглощения в области частот 63-8000 Гц для отделки стен и потолка помещений.

### 5.2.6 Повышенный уровень электромагнитных излучений

1) Источник возникновения фактора – дисплеи (мониторы). Они представляют собой источники наиболее вредных излучений, неблагоприятно влияющих на здоровье человека;

2) воздействие фактора на организм человека – при длительном воздействии данного фактора возникают жалобы на слабость, раздражительность, быструю утомляемость и ослабление памяти;

3) допустимые нормы: уровни электромагнитного поля приведены в таблице 24 согласно ГОСТ Р 50948-2001 “Средства отображения информации индивидуального пользования. Общие эргономические требования и требования безопасности.” [40].

Таблица 24 – Допустимые уровни электромагнитного поля

| **Наименование параметров** | | **ВДУ ЭМП** |
|---|---|---|
| Напряженность электрического поля | В диапазоне частот 5 Гц – 2 кГц | 25 В/м |
| | В диапазоне частот 2 кГц – 400 кГц | 2,5 В/м |
| Плотность магнитного потока | В диапазоне частот 5 Гц – 2 кГц | 250 нТл |
| | В диапазоне частот 2 кГц – 400 кГц | 25 нТл |
| Электростатический потенциал экрана видеомонитора | | 500 В |

4) предлагаемые средства защиты – рациональное размещение оборудования; использование средств, ограничивающих поступление

электромагнитной энергии на рабочие места персонала (экраны-фильтры и защитные очки).

### 5.2.7 Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека

В лаборатории при проведении исследований может возникнуть повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека

1) источник возникновения фактора – электрическое оборудование (микрокомпьютер, промышленный контроллер);

2) воздействие фактора на организм человека – электрический ток, проходя через организм, оказывает термическое, электролитическое и биологическое действие. Термическое действие выражается в ожогах отдельных участков тела, нагреве кровеносных сосудов, нервов и других тканей. Электролитическое действие выражается в разложении крови и других органических жидкостей, что вызывает значительные нарушения их физико-химических составов. Биологическое действие выражается в раздражении и возбуждении живых тканей организма, а также в нарушении внутренних биоэлектрических процессов, протекающих в нормально действующем организме и теснейшим образом связанных с его жизненными функциями;

3) допустимые нормы: согласно ГОСТ 12.1.038-82 ССБТ “Предельно допустимые значения напряжений прикосновений и токов” номинальное напряжение не превышает 50 В переменного тока (действующее значение) или 120 В постоянного (выпрямленного) тока. Напряжения прикосновения и токи, протекающие через тело человека при нормальном (неаварийном) режиме электроустановки, не должны превышать значений, указанных в таблице 25 [41];

Таблица 25 – Допустимые нормы.

| Род тока | U, В | I, мА |
|---|---|---|
| | Не более | |
| Переменный, 50 Гц | 2,0 | 0,3 |
| Переменный, 400 Гц | 3,0 | 0,4 |
| Постоянный | 8,0 | 1,0 |

4) предлагаемые средства – для защиты от поражения электрическим током все токоведущие части должны быть защищены от случайных прикосновений кожухами, корпус устройства должен быть заземлен. Заземление выполняется изолированным медным проводом сечением 1.5 мм, который присоединяется к общей шине заземления с общим сечением 48 мм. Общая шина присоединяется к заземлению, сопротивление которого не должно превышать 4 Ом.

### 5.2.8 Обоснование мероприятий по защите исследователя от действия опасных и вредных факторов

Для снижения влияния выявленных опасных и вредных факторов на работающих разработаны следующие мероприятия:

- Организация регулярных перерывов для сотрудников, работающих с данной библиотекой.
- Нормализация освещенности рабочих мест, путем установки дополнительных источников света и осветительных приборов.
- Все электроустановки должны быть снабжены средствами защиты, а также средствами оказания первой помощи в соответствии с действующими правилами применения и испытания средств защиты, используемых в электроустановках.

Применяемое в исследованиях электрооборудование, электротехнические изделия и материалы должны соответствовать требованиям государственных стандартов или технических условий, утвержденных в установленном порядке.

Согласно ГОСТ Р 12.1.019-2009 “Электробезопасность. Общие требования и номенклатура видов защиты” помещение, в котором

используется разрабатываемая библиотека, относится к классу помещений без повышенной опасности, в которых отсутствуют условия, создающие повышенную или особую опасность. В данных помещениях с установками напряжением до 1 кВ допускается применение неизолированных и изолированных токоведущих частей без защиты от прикосновения, если по местным условиям такая защита не является необходимой для каких-либо иных целей (например, для защиты от механических воздействий). При этом доступные прикосновению части должны располагаться так, чтобы нормальное обслуживание не было сопряжено с опасностью прикосновения к ним [42].

Таким образом, лаборатория, в которой проводятся исследования, соответствует «Правилам устройства электроустановок» и другим нормативам и не требует мероприятий по защите исследователя от действия опасных и вредных факторов.

## 5.3 Экологическая безопасность

В подразделе рассмотрен характер воздействия проектируемого решения на окружающую среду. Выявляются предполагаемые источники загрязнения окружающей среды, возникающие в результате разработки и реализации, предлагаемых решений.

### 5.3.1 Анализ влияния объекта и процесса исследования на окружающую среду

Объект исследования не оказывает влияния на окружающую среду, так как компьютер не осуществляет выбросов вредных веществ в атмосферу и гидросферу.

При завершении срока службы ПК их можно отнести к отходам электронной промышленности. Переработка таких отходов осуществляется разделением на однородные компоненты, химическим выделением пригодных для дальнейшего использования компонентов и направлением их для

дальнейшего использования согласно ГОСТ Р 55102-2012 “Ресурсосбережение. Обращение с отходами. Руководство по безопасному сбору, хранению, транспортированию и разборке отработавшего электротехнического и электронного оборудования, за исключением ртутьсодержащих устройств и приборов” [43]. Перечень элементов и содержащее их отработанное электротехническое и электронное оборудование, которые должны быть раздельно собраны при выводе отработавшего электротехнического и электронного оборудования из эксплуатации:

- Конденсаторы – содержат полихлорированные бифенилы;
- печатные платы и других устройств с площадью поверхности больше 10 см2 – содержат свинец, ртуть, кадмий;
- картриджи – содержат свинец, кадмий, бензол, толуол, фенол;
- пластик;
- электронно-лучевые трубки – содержат свинцовое стекло, соединения бария, люминофоры;
- элементы отработавшего электротехнического и электронного оборудования – содержат свинец, кадмий, олово;
- люминесцентные лампы – содержат ртуть.

### 5.3.2 Обоснование мероприятий по защите окружающей среды

Защита почвенного покрова и недр от твердых отходов реализуется за счет сбора, сортирования и утилизации отходов и их организованного захоронения. Главными нормативными актами, регулирующими вопрос утилизации персональных компьютеров, являются федеральные законы РФ «Об охране окружающей среды» и «Об отходах производства и потребления». Согласно этим законам, вся оргтехника подлежит утилизации с соблюдением определенных правил: демонтаж запчастей, сортировка отходов и утилизация.

Люминесцентные лампы относят к ртутьсодержащим отходам, и для их утилизации действует Постановление Правительства РФ [44].

Устанавливается порядок обращения с отходами производства и потребления в части осветительных устройств, электрических ламп, ненадлежащие сбор, накопление, использование, обезвреживание, транспортирование и размещение которых может повлечь причинение вреда жизни, здоровью граждан, вреда животным, растениям и окружающей среде.

Не допускается самостоятельное обезвреживание, использование, транспортирование и размещение отработанных ртутьсодержащих ламп потребителями отработанных ртутьсодержащих ламп, а также их накопление в местах, являющихся общим имуществом собственников помещений многоквартирного дома, за исключением размещения в местах первичного сбора и размещения и транспортирования до них. Сбор отработанных ртутьсодержащих ламп у потребителей осуществляют специализированные организации. Отходы, не подлежащие переработке и вторичному использованию, подлежат захоронению на полигонах.

## 5.4 Безопасность в чрезвычайных ситуациях

### 5.4.1 Анализ вероятных ЧС, которые может инициировать объект исследований

Объект исследований может инициировать возникновение такой чрезвычайной ситуации как пожар. Причинами пожара могут быть неисправность источника питания или компьютера.

### 5.4.2 Анализ вероятных ЧС, которые могут возникнуть в лаборатории при проведении исследований

При проведении исследований в лаборатории также может возникнуть пожар. Причинами пожара могут быть: игнорирование основных правил пожарной безопасности, неисправность электрической проводки, возгорание устройств искусственного освещения, возгорание устройств вычислительной аппаратуры вследствие нарушения изоляции или неисправности самой аппаратуры.

### 5.4.3 Обоснование мероприятий по предотвращению ЧС и разработка порядка действия в случае возникновения ЧС

Согласно НПБ 105-03 “Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности” помещение, в котором разрабатывалась система, относится к категории В3 по пожароопасности, содержит вещества и материалы, способные гореть при взаимодействии с водой, кислородом воздуха или друг с другом [45].

Помещение содержит ЭВМ, поэтому согласно СП 9.13130.2009 “Техника пожарная. ОГНЕТУШИТЕЛИ. Требования к эксплуатации” для ликвидации пожаров, вызванных возгоранием электрооборудования, применяются углекислотные огнетушители [46].

Для защиты от пожаров необходимо иметь в наличии такое пожарное оборудование как пожарные шкафы, пожарные щиты и огнетушители. Сотрудники должны уметь пользоваться таким оборудованием.

Сотрудники должны знать план эвакуации из помещения, расположение выходов из здания. Также необходимо проводить плановые эвакуации из здания, для того чтобы подготовить сотрудников к действиям в чрезвычайной ситуации.

Чтобы предотвратить пожар в производственном помещении, необходимо:

- работа должна проводиться только при исправном электрооборудовании;
- электросеть не должна перегружаться одновременно несколькими мощными потребителями электроэнергии;
- уходящий из помещения последним должен проверить выключены ли нагревательные приборы, электроприборы, оборудование и т.д.

При возникновении пожара тушить его самостоятельно целесообразно только на его ранней стадии при обнаружении загорания.

Согласно постановлению Правительства РФ от 16 сентября 2020 г. N 1479 "Об утверждении Правил противопожарного режима в Российской

Федерации" При обнаружении пожара или признаков горения (задымления, запаха гари, повышения температуры) в производственном помещении или на территории предприятия работник обязан немедленно сообщить об этом в пожарную охрану. Пожарной охране сообщается адрес объекта и место возникновения пожара. Сообщить пожарной охране необходимо даже в том случае, если загорание ликвидировано собственными силами. Огонь может остаться незамеченным в скрытых местах (в пустотах деревянных перекрытий и перегородок и т. д.), и впоследствии горение может возобновиться. Далее необходимо принять по возможности меры по эвакуации людей, тушению пожара и сохранности материальных ценностей [47].

## 5.5 Выводы

Таким образом, в данном разделе были рассмотрены вопросы соблюдения прав персонала на труд, выполнения требований к безопасности и гигиене труда, к промышленной безопасности, охране окружающей среды и ресурсосбережению.

Было выявлено, что разрабатываемая система не оказывает влияния на атмосферу и гидросферу, но для того, чтобы предотвратить загрязнение литосферы необходимо производить утилизацию отходов в ходе исследования.

Также в ходе работ было выяснено, что объект исследования или выполнение исследований в лаборатории могут инициировать возникновение такой чрезвычайной ситуации как пожар.

Помимо этого, были определены опасные и вредные факторы, которые могут возникнуть на разных этапах работ, а именно: перенапряжение анализаторов, в том числе вызванное информационной нагрузкой, отсутствие или недостаток естественного света, недостаточная освещенность рабочей зоны и повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека.

Полученными результатами также являются рекомендации по устранению и предотвращению пожаров, проведению утилизации отходов и устранения опасных и вредных факторов, рассматриваемых в данной работе. Произведены расчеты для создания освещенности E = 300 ЛК для лабораторного помещения, в котором изготовлялась библиотека. Также были сформулированы требования к организации рабочего места и организации труда. Данные рекомендации и требования могут быть внедрены на предприятиях, где планируется использование рассматриваемой библиотеки.

## Заключение

В ходе выполнения данной работы были выполнены следующие задачи:

1. Проведен обзор и сравнение основных синтаксических парсеров для русского языка на последней версии корпуса тестов syntagrus. Наилучшую точность показал парсер qbic, но из-за отсутствия подробной документации для разработки ПО использован парсер DeepPavlov.

2. Разработан модуль для извлечения временных конструкций. Данный модуль использует контекстно-грамматический парсер Spacy. Точность работы данного модуля составляет 95.5% на тестовом наборе данных. Также данный модуль способен выполнять нормализацию ВК, то есть преобразования даты к формату ГГГГ-ММ-ДД при помощи стандартной библиотеки datetime языка Python и библиотеки dateutil.

3. Разработан модуль для извлечения событий, связанных с ними, работающий с точностью 70%. Данный модуль на основе результатов синтаксического парсинга прокладывает маршрут в древе зависимостей от найденной ВК к связанному с ней событию.

4. Для решения возникшей проблемы пропущенных отрицательных частиц был разработан модуль поиска отрицаний, который работает с точностью 78,4%. Данный модуль на основе результатов синтаксического парсинга ищет связь в древе зависимостей от отрицательной частицы, найденной при помощи контекстно-грамматического парсинга, к связанной с ней сущности.

5. Добавлена обработка неопределенностей, объединяющая нечеткую логику и метод интервалов с расширяющимися границами.

Данная работа поможет переложить события из анамнеза на временную шкалу и собрать данные для последующей обработки и моделирования.

В дальнейшем могут быть произведены следующие улучшения системы:

1. Нормализация временных конструкций, которые зависят от других событий в анамнезе.
2. Добавление темпоральной логики.
3. Поддержка немедицинских данных.

**Список используемых источников**


1. Tang B. [и др.]. A hybrid system for temporal information extraction from clinical text // Journal of the American Medical Informatics Association. 2013. № 5 (20). С. 828–835.
2. Ning Q., Subramanian S., Roth D. An improved neural baseline for temporal relation extraction // EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. 2020. С. 6203–6209.
3. Lin C. [и др.]. Self-training improves Recurrent Neural Networks performance for Temporal Relation Extraction 2019. № Louhi. С. 165–176.
4. Goyal T., Durrett G. Embedding time expressions for deep temporal ordering models // ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. 2020. С. 4400–4406.
5. Galvan D. [и др.]. Investigating the Challenges of Temporal Relation Extraction from Clinical Text 2019. № 2016. С. 55–64.
6. Mirza P., Tonelli S. CATENA: CAusal and temporal relation extraction from natural language texts // COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers. 2016. С. 64–75.
7. Lee H. J. [и др.]. UTHealth at SemEval-2016 task 12: An end-to-end system for temporal information extraction from clinical notes // SemEval 2016 - 10th International Workshop on Semantic Evaluation, Proceedings. 2016. С. 1292–1297.
8. Liu S. [и др.]. Attention Neural Model for Temporal Relation Extraction 2019. С. 134–139.
9. Vashishtha S., Durme B. van, White A. S. Fine-grained temporal relation extraction // ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. 2020. С. 2906–2919.

10. Lin C. [и др.]. Representations of Time Expressions for Temporal Relation Extraction with Convolutional Neural Networks 2017. C. 322–327.

11. Lin C. [и др.]. A BERT-based Universal Model for Both Within- and Cross-sentence Clinical Temporal Relation Extraction // Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019. (2). C. 65–71.

12. Sohn Dr. S. [и др.]. Comprehensive temporal information detection from clinical text: Medical events, time, and TLINK identification // Journal of the American Medical Informatics Association. 2013. № 5 (20). C. 836–842.

13. Xu Y. [и др.]. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge // Journal of the American Medical Informatics Association. 2013. № 5 (20). C. 849–858.

14. dateparser – python parser for human readable dates — DateParser 0.7.6 documentation [Электронный ресурс]. URL: https://dateparser.readthedocs.io/en/latest/ (дата обращения: 01.10.2020).

15. orsinium-archive/rutimeparser: Recognize date and time in russian text and return datetime.datetime. [Электронный ресурс]. URL: https://github.com/orsinium-archive/rutimeparser (дата обращения: 01.10.2020).

16. Previous shared tasks | CoNLL [Электронный ресурс]. URL: https://www.conll.org/previous-tasks (дата обращения: 01.10.2020).

17. All treebanks [Электронный ресурс]. URL: https://universaldependencies.org/conll18/results-las.html (дата обращения: 01.10.2020).

18. UD_Russian-SynTagRus [Электронный ресурс]. URL: https://universaldependencies.org/treebanks/ru_syntagrus/index.html (дата обращения: 01.10.2020).

19. N L. O., Vinogradov Russian V. V, Ailamazyan A. K. GRAMEVAL 2020 SHARED TASK: RUSSIAN FULL MORPHOLOGY AND UNIVERSAL DEPENDENCIES PARSING 1 // Dialogue. 2020.

20. natasha/slovnet: Deep Learning based NLP modeling for Russian language [Электронный ресурс]. URL: https://github.com/natasha/slovnet (дата обращения: 01.10.2020).

21. Syntactic parsing — DeepPavlov 0.12.1 documentation [Электронный ресурс]. URL: http://docs.deeppavlov.ai/en/master/features/models/syntaxparser.html (дата обращения: 01.10.2020).

22. spaCy · Industrial-strength Natural Language Processing in Python [Электронный ресурс]. URL: https://spacy.io/ (дата обращения: 01.10.2020).

23. CoNLL 2018 Shared Task [Электронный ресурс]. URL: http://universaldependencies.org/conll18/evaluation.html (дата обращения: 01.10.2020).

24. natasha/yargy: Rule-based facts extraction for Russian language [Электронный ресурс]. URL: https://github.com/natasha/yargy (дата обращения: 01.10.2020).

25. Томита-парсер — Технологии Яндекса [Электронный ресурс]. URL: https://yandex.ru/dev/tomita/ (дата обращения: 26.03.2021).

26. Nagypál G., Motik B. A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2003. (2888). С. 906–923.

27. Dagum P., Luby M. Approximating probabilistic inference in Bayesian belief networks is NP-hard // Artificial Intelligence. 1993. № 1 (60). С. 141–153.

28. Hripcsak G. [и др.]. Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem // Journal of the American Medical Informatics Association. 2005. № 1 (12). С. 55–63.

29. Burney A. [и др.]. Conceptual fuzzy temporal relational model (FTRM) for patient data // WSEAS Transactions on Information Science and Applications. 2010. № 5 (7). С. 725–734

30. Funkner A. A., Kovalchuk S. V. Time expressions identification without human-labeled corpus for clinical text mining in russian Springer, 2020.С. 591–602.

31. СП 2.4.3648-20 Санитарно-эпидемиологические требования к организациям воспитания и обучения, отдыха и оздоровления детей и молодежи.

32. Трудовой кодекс Российской Федерации от 30.12. 2001 г. № 197–ФЗ (ред. от 01.04.2019 г.). – М., 2015. – 123 с.

33. Федеральный закон от 27.07.2006 N 152-ФЗ (ред. от 30.12.2020) “О персональных данных.”

34. ГОСТ Р 50923-96 Дисплеи. Рабочее место оператора.

35. ГОСТ 12.0.003-2015 Система стандартов безопасности труда (ССБТ). Опасные и вредные производственные факторы. Классификация.

36. СП 52.13330.2016 Естественное и искусственное освещение.

37. СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания.

38. ГОСТ 6825-91 Лампы люминесцентные трубчатые для общего освещения.

39. ГОСТ Р ИСО 9241-5-2009 Эргономические требования к проведению офисных работ с использованием видеодисплейных терминалов.

40. ГОСТ Р 50948-2001 Средства отображения информации индивидуального пользования. Общие эргономические требования и требования безопасности.

41. ГОСТ 12.1.038-82 ССБТ Предельно допустимые значения напряжений прикосновений и токов.

42. ГОСТ Р 12.1.019-2009 Электробезопасность. Общие требования и номенклатура видов защиты.

43. ГОСТ 55102-2012 Ресурсосбережение. Обращение с отходами. Руководство по безопасному сбору, хранению, транспортированию и разборке

отработавшего электротехнического и электронного оборудования, за исключением ртутьсодержащих устройств и приборов.

44. Постановление Правительства РФ от 03.09.2010 № 681 (ред. от 01.10.2013) Об утверждении Правил обращения с отходами производства и потребления в части осветительных устройств, электрических ламп, ненадлежащие сбор, накопление, использование, обезвреживание.

45. НПБ 105-03 Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности.

46. СП 9.13130.2009 Техника пожарная. ОГНЕТУШИТЕЛИ. Требования к эксплуатации.

47. Постановление Правительства РФ от 16 сентября 2020 г. № 1479 Об утверждении Правил противопожарного режима в Российской Федерации.

## ПРИЛОЖЕНИЕ А

(справочное)

**Algorithms and software for time expression identification in semi-structured electronic medical texts**

Студент

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8ПМ9И | Журман Дмитрий Александрович | | |

Руководитель ВКР

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент | Аксёнов Сергей Владимирович | Кандидат технических наук | | |

Консультант-лингвист отделения иностранных языков ШБИП

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент | Сидоренко Татьяна Валерьевна | Кандидат педагогических наук | | |

## Introduction

During staying in a hospital when being ill, various events might occur with a patient. For example, he/she receives a treatment that causes new symptoms which lead to surgeries, etc. If we transfer all events from anamnesis to a timeline, a doctor will be able to see clearly what has happened to the patient and when. The timeline will also prepare a dataset for the predictive models in medicine and healthcare. The information about these events can be collected from the anamnesis of Electronic Medical Records (EMR). Unfortunately, the patient's anamnesis is unstructured (free texts in natural language). Also, there is no labelled corpus of medical texts in Russian language. Therefore, the solution to this problem is possible via syntactical and morphological analysis or rule-based methods.

Thus, the goal of this study is to develop some methods to identify time expressions (TE) and their corresponding events. At the same time, the proposed methods should work with an unlabelled input text corpus.

To achieve the goal, the following tasks are to be fulfilled:

1. review and compare syntactic parsers for the Russian language;

2. develop a module for time expressions identification;

3. develop a module for searching for the events associated with time expressions.

## Systems for extracting time expressions from medical texts

At the moment, there have been four competitions for the extraction of temporary structures I2B2 Temporal Challenge (2013), Clinical TempEval Shared Task (2015, 2016, and 2017). In all competitions, it was required to find events, temporary structures, and relations between them using annotated corpus of texts [1].

Solutions based on Clinical TempEval Shared Task used supervised learning methods to solve the problem. The most commonly used Recurrent Neural Networks (LSTM / GRU) [2–5]. Earlier articles used SVM models [6,7], and more modern works use attention models [8,9]. In some cases, researchers applied Convolutional Neural Networks [10] or BERT model [11] for Temporal Relation Extraction. In any

case, we cannot use the methods described in these works to solving the problem, since they used a labeled corpus for English for training, and there is no labeled corpus of medical texts for the Russian language.

In the I2B2 Temporal Challenge, participants used combined methods, i.e. supervised learning and rule-based methods. Rule-based methods have been used to find temporal constructs and relations between events. For example, the test results always come before the diagnosis. Also, sometimes syntactical parsing was applied to find candidates for events. Finally, events were extracted with help of the SVM model [1,12].

From the above overview of the methods applied in the I2B2 Temporal Challenge, it follows:

1. For the extraction of time expressions with an F-score of 0.85 and higher can be achieved using methods based on grammar rules.

2. As potential events associated with time expressions, you can use any phrase that has a connection with expression based on the parsing of sentence dependencies.

These approaches will be used in the further development of algorithms and software for the identification of time expressions, in semi-structured electronic medical texts.

**Designing software for identification of time expressions**

Figure 1 shows three main parts of the method. To achieve the purpose, first of all, it is necessary to implement the extraction of time expressions from sentences using rule-based methods. These time expressions should be normalized, that is, brought to a single format (YYYY-MM-DD). At the moment, there are already a number of ready-made libraries (dateparser and rutimeparser [15]) for extracting time expressions and normalizing them. However, they show low accuracy when working with medical text. This is due to the fact that the corpus used in this work contains repetitive events and has many formats of time expressions. Therefore, we developed a module that used rule-based methods for extracting time expressions from sentences.

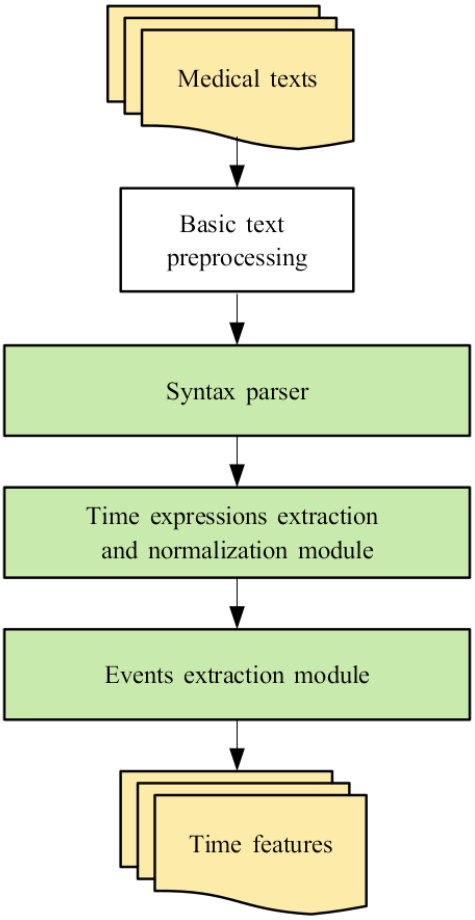


Figure 1 – General scheme of the method

Then we need to parse sentences containing time expressions with one of the syntactical parsers. After that, in the syntactic tree, we need to build a path from the found time expression to the event to which it belongs. We created the algorithm with recursion shown in Figure 2 to extract events. If a complex sentence is input to the module, the module splits it into parts. If time expression in the first part of the sentence, then the algorithm works with this part as with an ordinary simple sentence, ignoring other parts. If time expression is not in the first part, the module removes the first part from the proposal and re-determines where the time expression is. This continues until time expression is in the first part of the sentence or only 1 part remains.

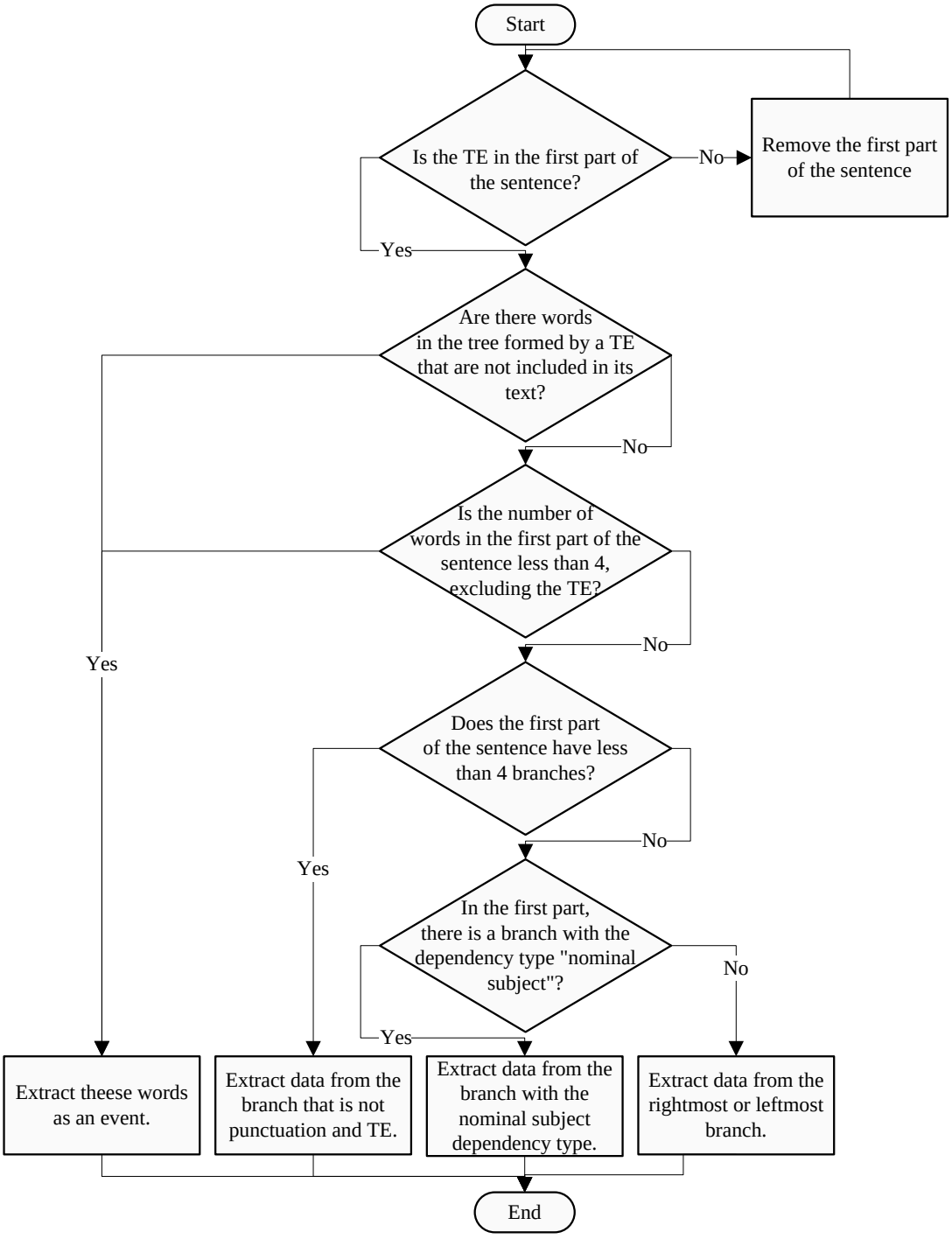


Figure 2 – The algorithm of the module for extracting events

When parsing a simple sentence, the module first checks if there are any words in the tree formed by the time expressions that are not included in its text. If there are many branches in the first part of the sentence and all of them are not deep, then the module extracts the words included in this branch as events. If there are less than four branches in a sentence, then usually the first branch is punctuation, the second is time expression, and therefore the 3rd is the event corresponding to it. If there are more than 3 branches in the sentence part, and they are deep, then the

module starts looking at the dependency types. Most often, an event has the nominal subject type of dependency. If there is a branch in the tree with this connection, then the words in this branch are an event. If a part of the sentence does not satisfy any of the conditions, then data is extracted from the rightmost or leftmost branch.

For example, in the sentence "Constantly worried about dizziness, in the summer of 2010 a single episode of loss of consciousness." (Постоянно беспокоит головокружение, летом 2010 года однократно эпизод потери сознания), the module, after failing to find the time expression in the first part, remove it, and then works with the part “in the summer of 2010 once episode of loss of consciousness” as with an ordinary simple sentence. Figure 3 shows result of syntactical parsing for this sentence. In the second part of sentence there is no words in the tree formed by time expression, there are more than 4 words and there are 4 branches and there is no branch with the dependency type “nominal subject”. It means that algorithm extracted data from the rightmost branch and the result is “episode of loss of consciousness” (эпизод потери сознания).

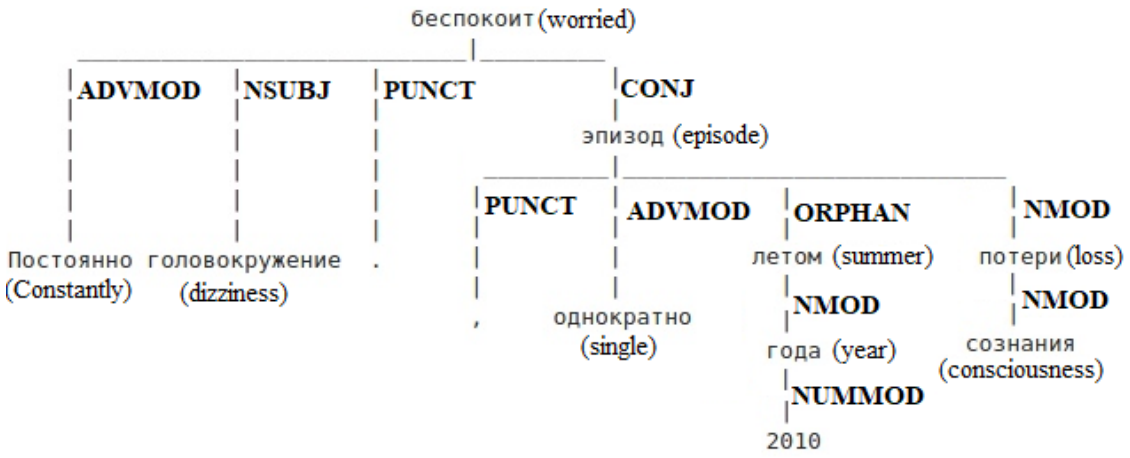


Figure 3 – Syntactical tree of sentence. Some prepositions and articles have been omitted due to their absence in Russian.

**Choosing a syntactical parser for the Russian language**

Developing syntactical parser requires a huge corpus of texts labeled by professional linguists and significant computing resources. As mentioned above, labeled corpus of medical texts for the Russian language is not publicly available. Therefore, in this work, we will use a ready-made syntactical parser for the Russian language. Therefore, this section provides a comparison of existing syntactic parsers.

CoNLL Shared Task is an annual competition among syntactic parsers. However, in 2019, the competition was held only for English, and in 2020 the competition has not yet taken place [16]. Table 1 shows the result of CoNLL Shared Task 2018 [17]. The CoNLL Shared Task compares how parsers work with the Russian language in the syntagrus corpus. This corpus contains 61889 sentences and 1106296 tokens [18]. All sentences were labeled by linguists in the Universal Dependencies format.

In addition to the CoNLL Shared Task, there is the GramEval-2020 competition on complete syntactic parsing of texts in Russian. Table 2 shows the results of these competition [19].

These competitions used two types of metrics to evaluate the effectiveness of syntactical parsing: unlabeled attachment score (UAS) and labeled attachment score (LAS). LAS is a standard evaluation metric in dependency parsing: the percentage of words that are assigned both the correct syntactic head and the correct dependency label. UAS is the percentage of words that get the correct head [23].

Table 1 – The result of CoNLL Shared Task 2018.

| Parser | LAS |
|---|---|
| HIT-SCIR (Harbin) | 92.48 |
| TurkuNLP (Turku) | 91.72 |
| Stanford (Stanford) | 91.59 |
| UDPipe Future (Praha) | 91.46 |
| ParisNLP (Paris) | 91.41 |

Table 2 – The result of GramEval-2020.

| | Fiction | News | Poetry | Social | Wiki | 17 cent |
|---|---|---|---|---|---|---|
| qbic | 0.896 | 0.912 | 0.814 | 0.807 | 0.781 | 0.665 |
| ADVance | 0.869 | 0.911 | 0.780 | 0.784 | 0.760 | 0.618 |
| lima | 0.850 | 0.843 | 0.725 | 0.713 | 0.697 | 0.546 |
| vocative | 0.826 | 0.834 | 0.660 | 0.659 | 0.694 | 0.500 |
| Turku | 0.859 | 0.877 | 0.731 | 0.733 | 0.711 | 0.502 |
| Stanford | 0.854 | 0.873 | 0.709 | 0.706 | 0.703 | 0.509 |
| UDPipe | 0.811 | 0.817 | 0.666 | 0.644 | 0.668 | 0.462 |
| SyntaxNet | 0.808 | 0.802 | 0.6 | 0.614 | 0.645 | 0.446 |
| MaltParser | 0.599 | 0.553 | 0.404 | 0.476 | 0.436 | 0.340 |

Table 3 shows the comparison results. The qbic solution showed the best score in terms of the LAS metric, despite the fact that it was not trained on the

syntagrus corpus. It is also worth noting the good score of the DeepPavlov parser. However, the qbic parser is a solution for specific competitions GramEval-2020, and not a ready-made package for any task, unlike Stanza, UDPipe, DeepPavlov, and Slovnet. Repositories of such parsers from the CoNLL 2018 Shared Task competitions show that these solutions have not been updated since the end of the competition. Therefore, in this research, we will use the DeepPavlov syntactic parser, since it is regularly updated and contains detailed documentation.

Table 3 – The result of parser comparison

| Parser | Speed (sent/s) | UAS | LAS | Sensitivity to puntcuation |
|---|---|---|---|---|
| qbic | 55 (GPU) | 96.98 | 94.16 | - |
| DeepPavlov | 53 (GPU) | 93.40 | 92.07 | - |
| Stanza | 18 (GPU) | 91.75 | 89.97 | + |
| Turku | 21 (CPU) | 88.36 | 87.08 | + |
| UDPipe | 87 (GPU) | 87.67 | 80.06 | + |
| Slovnet | 705 (CPU) | 81.05 | 76.43 | - |
| Spacy | 40 (CPU) | 86.93 | 75.56 | + |

**Development of a module for extraction time expressions**

Among the existing open-source libraries in for writing rules for the Russian language, we can single out Spacy and Yargy [22,24]. Initially, all 110 rules (each of the rules corresponds to 1–3-time expressions) were written in yargy, but the processing speed was slow (7 sentences per second). After the rules were rewritten for spacy, it was possible to achieve a speed of about 35 sentences per second.

In addition to extracting a temporary construction, the module is able to normalize it and determine its type (long-term, one-time, repetitive, dependent). For normalization, we used ready-made libraries dateparser and rutimeparser, which are able to normalize the most common time expressions (in 2009, 03.12.2010, 3 years ago), for specific time expressions (from 2009 to 2010, from 51 years, summer 2010), we made normalization rules or realized specific preprocessing methods for dateparser. During normalization, there was a problem with dependent events (after a month), since they depend on previous events in the anamnesis (for example, after

discharge a month later, she noted the resumption of paroxysms of arrhythmia.). Table 4 shows the examples of work of the module for extracting time expressions.

In this research, we used a set of anonymized EMRs of patients with the acute coronary syndrome (ACS) who admitted to Almazov National Medical Research Centre (2000 sentences) and Tomsk NRMC Cardiology Research Institute (7777 sentences). Finally, to measure the accuracy of the solution, it is necessary to label a data set that was not used in the development of a method for solving the problem.

To measure the accuracy, a dataset of 500 sentences was labeled. On this dataset, the accuracy of extracting temporary structures was 91.60%, and the accuracy of normalization was 80.40%.

Table 4 – The examples of work of the module for extracting time expressions.

| Sentence in English (in Russian) | Time expression | Normal form | Stamp |
|---|---|---|---|
| In the fall of 2010, the patient was hospitalized routinely in the endocrinology department. (Осенью 2010 года пациентка госпитализировалась в плановом порядке в отделение эндокринологии.) | In the fall of 2010 (Осенью 2010 года) | 01.09.2010 | one-time |
| From April 2010 on sick leave. (С апреля 2010 на больничном.) | From April 2010 (С апреля 2010) | 01.04.2010 | long-term |
| He takes warfarin 2 times a day. (2 раза в день принимает варфарин.) | 2 times a day (2 раза в день) | - | repetitive |
| After discharge a month later, she noted the resumption of paroxysms of arrhythmia. (После выписки через месяц отметила возобновление пароксизмов аритмии.) | a month later (через месяц) | - | dependent |
| From the age of 50, pressing chest pains. (С 50 лет давящие боли в груди.) (birthday 14.03.1942) | From the age of 50 (С 50 лет) | 14.03.1992 | long-term |

**Development of a module for events extraction**

Module for events extraction was developed using the spacy library as it contains tools for navigating the syntax tree. After that, we evaluate the accuracy of the test dataset from the previous section. When evaluated using the accuracy metric, the score was 53.80%. However, this metric is sensitive to word order and

punctuation; therefore, Table 5 shows a detailed analysis of the errors. From the analysis, it can be seen that the most common error for the accuracy metric was “controversial labeling”. This is due to the fact that the test dataset was not performed ideally, that is, sometimes two events in a sentence refer to the same time expressions. Thus, the module extracts one event, and another was included in the test dataset. Also, sometimes the module extracts an addition on in the event text that was not included in the test dataset, or, on the contrary, does not find the addition that was in the test dataset. Thus, all 97 sentences in this category are not really errors. Apart from the different word order and controversial labeling, the event extraction accuracy is 74.80%.

Table 5 – Analysis of the errors for 500 labeled sentences.

| Error | Amount |
|---|---|
| Controversial labeling | 97 |
| Syntactic parser error | 28 |
| Error extracting event | 26 |
| Error extracting time expression | 28 |
| Not all homogeneous terms found | 13 |
| Different word order | 9 |
| Missing preposition | 9 |
| Spelling mistake in a sentence | 8 |
| Word "NOT" is missing | 5 |
| In sentence there are two events related to two different time expression in one branch | 4 |
| Error extracting events and incorrectly composed sentence | 3 |
| Dash missed | 2 |
| Controversial labeling, incorrectly composed sentence | 1 |
| A grammatical and spelling mistake in a sentence | 1 |
| Not all words related to the event were extracted | 1 |
| Syntactic parser error and incorrectly composed sentence | 1 |
| Part of the event is in the subordinate clause | 1 |
| A grammatical error in a sentence | 1 |
| Syntax parser error and error extracting time expression | 1 |
| Complex grammatical structure | 1 |
| Syntactic parser error due to the presence of words in Latin | 1 |

**Development of a module for negations extraction**

In some cases, the parser makes a serious mistake and extracts the event correctly, but without a NOT particle or a WITHOUT particle, which radically changes the meaning. This is because in the syntactic tree the particles NOT and WITHOUT are sometimes located in a different branch from the event, and in the

Russian language there is no special type of connection for them, unlike the English language. It is necessary not only to correctly define the negative particle but also to find the negated expression. This will make it possible to understand whether negation enters time expression or not. Solving this problem will help to increase the accuracy of the module. It will also allow the development of an additional tool for processing medical documents.

To fix this error, syntactic parsing was used. The Spacy parser was used to find the most common negation patterns.

A sentence with negation is sent to the input of the module. If the negative particle is the words "deny", "refuse" or "no", words from the branch with the link "nominal subject" are extracted from the tree formed by the negative particle.

After that, there is a check if there is a connection with the "nominal subject" type in the entire tree. In this case, all words from the given branch are extracted.

If the sentence is complex, then the module determines the part containing the negative particle and continues to work with it as with a simple sentence.

If the top of the tree is a noun, then the top, the child of the top, and the negative particle are extracted.

If the tree has fewer than four branches, then all words are extracted except punctuation. If none of the checks passed, then only the top of the tree and the negative particle are extracted.

After checking the module on a labeled dataset, an accuracy of 78.4% was achieved. Embedding this negation search algorithm into the software will help to more accurately determine the events associated with time expressions.

**Handling fuzziness**

One of the problems arising when working with time expressions is fuzziness. For example, when the expression "10 years ago" appears in a patient's chart, the point is not meant exactly 10 years ago. The necessary event can be in a certain range around this point.

In the temporal data model, the following types of fuzziness can be distinguished [26]:

- uncertainty. Some of the temporal information is unknown or imprecise. For example, the date of birth of Socrates is "Type 2 diabetes was diagnosed about 5 years ago.";
- subjectivity. Temporal events or periods can be subjectively or ambiguously defined. For example, " There were no heart attacks after the operation.";
- vagueness. Events can be defined with different granularities or vagueness. For example, “There are no episodes of sleepiness during the day.”.

It might be possible to model vagueness "10 years ago" as a probability density function over time with the greatest probability around 10 years (For example, gaussian function), but calculating conclusions from such a representation is NP-hard [27].

Approach performed in [28], model vagueness by widening the limits of the constraints. This method represented an assertion as plus or minus one unit with a minimum of one-half unit. For example, “The rash appeared three weeks before admission” was represented as a range from two to four weeks. Vagueness modifiers such as “about” and “around” can widen the interval: from −50% to +100%. “About three weeks” can be represented as 1.5 to six weeks. The size of the range may depend on the context.

Another approach to represent vagueness is fuzzy logic. In fuzzy set theory, if we want to represent a crisp interval *i* when an event happens, we can use membership function *I*, which represents our confidence level that *t* is in *i*. If *I(t) = 0*, we are completely confident that *t* is not in *i*; if *I(t) = 1*, we are completely confident that *t* is in *I* [26]. A membership function can be designed in several ways [29].

Triangular membership functions:

$$f(x,a,b,c) = \max(\min(\frac{x-a}{b-a}, 1, \frac{c-x}{c-b}), 0) \quad (1)$$

Trapezoidal membership function:

$$f(x,a,b,c,d) = \max(\min(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}), 0) \quad (2)$$

Figure 4 shows Triangular and Trapezoidal membership functions.

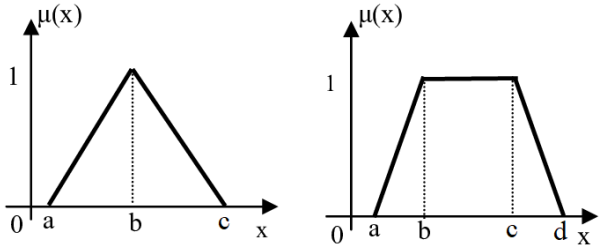


Figure 4 – Triangular membership and Trapezoidal membership functions

To handle fuzziness in this work, we applied an approach that combines fuzzy logic and the method of intervals with expanding boundaries.

If a one-time time expression is found (for example, a month ago), then it is represented as a triangular membership function centered at a point exactly one month ago. The boundaries of the membership function are defined as plus or minus one time unit (day, week, month, or year) from the point on the timeline to which this event belongs.

If a long-term time expression is found (for example, the last 2 months), then it is represented as a trapezoidal membership function. Points *b* and *c* in Figure 5 correspond to the boundaries of the normalized time expression. The boundaries of the membership function are defined as ±one time unit (day, week, month, or year) from the boundary of the normalized temporal construction. If point *c* corresponds to the present moment in time, then point *d* is also equal to the present moment in time.

If the words 'about', 'approximately', 'approximately', 'almost' or 'somewhere' are contained in the text of time expression, then the width of the membership function increases by 2 times.

Also, when a patient tells a doctor that an event happened to him 5 years ago, he is more confident in this than when he says that an event happened to him 20 years ago. In the second case, the actual date of this event may be outside the interval from 19 years ago to 21 years ago. To solve this problem, a coefficient was added that increases the width of the membership function for events that occurred more than 5 years ago. This coefficient is calculated using the following formula:

$$k = 1 + \frac{past - 43800}{175200} \quad (3)$$

where past – the amount of time elapsed from the current time in hours;

43800 – number of hours in 5 years;

175200 – number of hours in 20 years.

**Testing the developed software**

When writing rules for extracting temporary structures, creating an event extraction module, and measure accuracy, we used the data set of the Research Institute of Cardiology of Tomsk (7777 sentences). This data set contained pre-selected sentences with time expressions.

To measure the accuracy, we made an experiment on the data set of the Almazov National Medical Research Centre (2000 proposals), which was not used in writing the rules and creating the algorithm. Unlike the previous data set, it contained sentences with and without time expressions. On this dataset, the module sometimes extracted phrases that did not contain time expressions. To solve this problem, we used a pre-trained classifier, which determines whether the sentence includes a time expression. After applying this classifier, we found that 580 sentences contain time expression. Then a module was applied to extract events. Table 6 shows the analysis of the module errors. Thus, the accuracy of event extraction on the new dataset was 70%.

Table 6 – Analysis of the errors on the dataset of the Almazov Center

| Error | Amount |
|---|---|
| Error extracting event | 77 |
| Error extracting time expression | 36 |
| Syntactic parser error | 31 |
| Extract unnecessary words | 9 |
| Spelling mistake in a sentence | 6 |
| Missing preposition | 6 |
| Syntactic parser error and error extracting time expression | 3 |
| A grammatical error in a sentence | 2 |
| Complex grammatical structure | 1 |
| Complex grammatical structure with grammatical and spelling mistakes in a sentence | 1 |
| Complex grammatical structure with spelling mistakes in a sentence | 1 |
| Not all homogeneous terms found | 1 |

## Conclusion

According to the results of the experiment, the most common causes of errors are the algorithm of the event extraction module. Since in this study we are dealing with a natural unstructured language, it is impossible to develop an algorithm that would fit all sentences written in free text. When analyzing the errors, we found sentences that contradict the algorithm and their correction leads to even more errors.

We can see errors when extracting time expression on a new dataset because there are three new forms of time expression that appeared in the test dataset, and grammar rules are not written for them.

In some cases, errors occurred in the syntactic parser, as mentioned above, due to the absence of a labeled corpus of medical texts, it is impossible to do something with this category of errors.

In addition, we found cases when the module does not find all homogeneous terms, taking them for a subordinate clause.

Thus, we propose a module for extracting time expressions and a module for extracting events associated with them, working with an accuracy of 70%. Also, this module able to normalize extracted expressions. This work will help to transfer events from the anamnesis to the timeline and collect data for further processing and modeling. In the future, we plan to do the normalization of temporary structures, which depend on other events in the anamnesis and add temporal logic.

**References**

1. Tang B. et al. A hybrid system for temporal information extraction from clinical text // J. Am. Med. Informatics Assoc. 2013. Vol. 20, № 5. P. 828–835.

2. Ning Q., Subramanian S., Roth D. An improved neural baseline for temporal relation extraction // EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf. 2020. P. 6203–6209.

3. Lin C. et al. Self-training improves Recurrent Neural Networks performance for Temporal Relation Extraction. 2019. № Louhi. P. 165–176.

4. Goyal T., Durrett G. Embedding time expressions for deep temporal ordering models // ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. 2020. P. 4400–4406.

5. Galvan D. et al. Investigating the Challenges of Temporal Relation Extraction from Clinical Text. 2019. № 2016. P. 55–64.

6. Mirza P., Tonelli S. CATENA: CAusal and temporal relation extraction from natural language texts // COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap. 2016. P. 64–75.

7. Lee H.J. et al. UTHealth at SemEval-2016 task 12: An end-to-end system for temporal information extraction from clinical notes // SemEval 2016 - 10th Int. Work. Semant. Eval. Proc. 2016. P. 1292–1297.

8. Liu S. et al. Attention Neural Model for Temporal Relation Extraction. 2019. P. 134–139.

9. Vashishtha S., van Durme B., White A.S. Fine-grained temporal relation extraction // ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. 2020. P. 2906–2919.

10. Lin C. et al. Representations of Time Expressions for Temporal Relation Extraction with Convolutional Neural Networks. 2017. P. 322–327.

11. Lin C. et al. A BERT-based Universal Model for Both Within- and Cross-sentence Clinical Temporal Relation Extraction // Proc. 2nd Clin. Nat. Lang.

Process. Work. 2019. Vol. 2. P. 65–71.

12. Sohn Dr. S. et al. Comprehensive temporal information detection from clinical text: Medical events, time, and TLINK identification // J. Am. Med. Informatics Assoc. 2013. Vol. 20, № 5. P. 836–842.

13. Xu Y. et al. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge // J. Am. Med. Informatics Assoc. 2013. Vol. 20, № 5. P. 849–858.

14. dateparser – python parser for human readable dates — DateParser 0.7.6 documentation [Electronic resource]. URL: https://dateparser.readthedocs.io/en/latest/ (accessed: 01.10.2020).

15. orsinium-archive/rutimeparser: Recognize date and time in russian text and return datetime.datetime. [Electronic resource]. URL: https://github.com/orsinium-archive/rutimeparser (accessed: 01.10.2020).

16. Previous shared tasks | CoNLL [Electronic resource]. URL: https://www.conll.org/previous-tasks (accessed: 01.10.2020).

17. All treebanks [Electronic resource]. URL: https://universaldependencies.org/conll18/results-las.html (accessed: 01.10.2020).

18. UD_Russian-SynTagRus [Electronic resource]. URL: https://universaldependencies.org/treebanks/ru_syntagrus/index.html (accessed: 01.10.2020).

19. N L.O., Vinogradov Russian V. V, Ailamazyan A.K. GRAMEVAL 2020 SHARED TASK: RUSSIAN FULL MORPHOLOGY AND UNIVERSAL DEPENDENCIES PARSING 1 // Dialogue Can. Philos. Assoc. 2020.

20. natasha/slovnet: Deep Learning based NLP modeling for Russian language [Electronic resource]. URL: https://github.com/natasha/slovnet (accessed: 01.10.2020).

21. Syntactic parsing — DeepPavlov 0.12.1 documentation [Electronic resource]. URL: http://docs.deeppavlov.ai/en/master/features/models/syntaxparser.html (accessed: 01.10.2020).

22. spaCy · Industrial-strength Natural Language Processing in Python [Electronic resource]. URL: https://spacy.io/ (accessed: 01.10.2020).

23. CoNLL 2018 Shared Task [Electronic resource]. URL: http://universaldependencies.org/conll18/evaluation.html (accessed: 01.10.2020).

24. natasha/yargy: Rule-based facts extraction for Russian language [Electronic resource]. URL: https://github.com/natasha/yargy (accessed: 01.10.2020).

25. Томита-парсер — Технологии Яндекса [Electronic resource]. URL: https://yandex.ru/dev/tomita/ (accessed: 26.03.2021).

26. Nagypál G., Motik B. A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2003. Vol. 2888. P. 906–923.

27. Dagum P., Luby M. Approximating probabilistic inference in Bayesian belief networks is NP-hard // Artif. Intell. 1993. Vol. 60, № 1. P. 141–153.

28. Hripcsak G. et al. Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem // J. Am. Med. Informatics Assoc. 2005. Vol. 12, № 1. P. 55–63.

29. Burney A. et al. Conceptual fuzzy temporal relational model (FTRM) for patient data // WSEAS Trans. Inf. Sci. Appl. 2010. Vol. 7, № 5. P. 725–734.

30. Funkner A.A., Kovalchuk S. V. Time expressions identification without human-labeled corpus for clinical text mining in russian // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, 2020. Vol. 12140 LNCS. P. 591–602.

## ПРИЛОЖЕНИЕ Б

(обязательное)

### Диаграмма Ганта

| № работ | Вид работ | Исполнители | $T_{кi}$, кал. дн. | Продолжительность выполнения работ | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | февр. | | март | | | апр. | | | май | | | июнь | |
| | | | | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 |
| 1 | Постановка задачи | Руководитель | 3 | | | | | | | | | | | | | |
| | | Инженер | 4 | | | | | | | | | | | | | |
| 2 | Подбор и изучение материалов по теме | Инженер | 10 | | | | | | | | | | | | | |
| 3 | Разработка и утверждения ТЗ | Руководитель | 2 | | | | | | | | | | | | | |
| | | Инженер | 12 | | | | | | | | | | | | | |
| 4 | Календарное планирование работ по теме | Инженер | 2 | | | | | | | | | | | | | |
| 5 | Разработка вариантов исполнения проекта | Руководитель | 5 | | | | | | | | | | | | | |
| | | Инженер | 7 | | | | | | | | | | | | | |
| 6 | Подготовка данных | Инженер | 15 | | | | | | | | | | | | | |
| 7 | Подбор модели синтаксического парсера для русского языка | Инженер | 15 | | | | | | | | | | | | | |
| 8 | Разработка модуля для извлечения временных конструкций | Инженер | 15 | | | | | | | | | | | | | |
| 9 | Разработка модуля для поиска событий, связанных с временными конструкциями | Инженер | 15 | | | | | | | | | | | | | |
| 10 | Проверка работы библиотеки на тестовых данных и правилах | Руководитель | 2 | | | | | | | | | | | | | |
| | | Инженер | 8 | | | | | | | | | | | | | |
| 11 | Составление документации | Инженер | 9 | | | | | | | | | | | | | |

– руководитель  – инженер

# ПРИЛОЖЕНИЕ В

(обязательное)

## Реестр рисков

| № | Риск | Потенциальное воздействие | Вероятность наступления | Влияние риска | Уровень риска | Способы смягчения риска | Условия наступления |
|---|---|---|---|---|---|---|---|
| 1 | Несоответствие разработанной и требуемой функциональности | Невостребованность разработки | 2 | 3 | средний | Прототипирование, разработка сценариев использования, участие потенциальных пользователей | Ошибки при постановке задачи, недостаточный анализ качества разработки и ее перспективности на рынке |
| 2 | Постоянный поток изменений требований | Задержки выполнения работ | 2 | 2 | низкий | Установка ограничений для внесения изменений, итеративность | Ошибки при постановке задачи |
| 3 | Технологическое отставание | Невостребованность разработки | 2 | 2 | низкий | Технический анализ, анализ стоимости, прототипирование | Не достаточная оценка существующих моделей |
| 4 | Недостаточная производительность | Невостребованность разработки | 1 | 3 | средний | Проведение сравнительного тестирования, прототипирование | Ошибки при постановке задачи, недостаточный анализ качества разработки и ее перспективности |